THE PROCEEDINGS OF
THE PHYSICAL SOCIETY
Section B

VOL. 63, PART 11 1 November 1950 No. 371 B

A Note on the Orientated Overgrowths of Metal Films on Single Crystal Inorganic Substrates

BY H. R. THIRSK
Department of Physical Chemistry, King's College, University of Durham,
Newcastle-on-Tyne

MS. received 24*th April* 1950

ABSTRACT. Some new examples of metal crystal orientations on inorganic substrates are described. They include silver and iron on mica, platinum on the cube face of potassium chloride, iron and silver on the (111) face of potassium chloride, silver and nickel on the cube face of magnesium oxide. Some of the problems involved in the study of the types of crystal growth are discussed briefly for the cases examined.

§ 1. INTRODUCTION

EXAMPLES of the structures of thin evaporated films formed on crystalline substrates have been described by a number of workers. In a recent paper by van der Merwe (1949) concerned more particularly with the nature of the orientating forces at the metal substrate interface, a useful table was given summarizing a great amount of experimental data on epitaxial growth. In view of the value of such information, particularly as a source of data for theoretical studies, it has been thought worth while to describe in this note further examples from an electron diffraction investigation supplementing the data recorded there.

§ 2. EXPERIMENTAL

The following technique was used in the preparation of the evaporated metal films. The ends of a silica tube, 4 cm. in diameter and about 40 cm. long, were fixed by means of Apiezon wax into brass annuli with ground faces against which two brass end caps could be joined using Apiezon M grease on flat metal to metal faces. The leads for a filament, acting as a source of metal, and a thermocouple, registering the temperature of the substrate, were taken through one cap. The tube was evacuated through a wide tube in the second brass cap by means of a mercury diffusion pump backed by a Hyvac rotary oil pump. A mercury vapour trap was placed between the silica tube and the diffusion pump. The substrate on which the deposit was to be formed was heated by means of a furnace surrounding the silica tube.

Silver, nickel, platinum and iron deposits were obtained by evaporation from a filament of the pure metal. The filament was first flashed *in vacuo* and the silica tube heated over a suitable length to drive off adsorbed air. The degassed substrate, having been brought into position near the filament, was raised to a desired temperature by localized heating, after which the filament was heated electrically until a deposit of metal could be observed forming slowly on the substrate.

The following single crystal substrates were used during the work. Magnesium oxide in the form of single crystals with freshly cleaved cubic faces about 1 cm^2 in area; (111) faces of potassium chloride prepared by grinding the required face on a small cube of the substance, polishing on filter paper with water and finally very lightly etching with water; cube faces of potassium chloride prepared by cleavage; the cleavage face of calcite and the cleavage face of muscovite mica. Electron diffraction and microscope examination of (111) surfaces of KCl indicated that, as prepared, they were extremely smooth. All the electron diffraction patterns were taken on a Finch-type instrument with an accelerating voltage of approximately 50–55 kv.

Silver on Mica

Temperatures below 200° c. gave silver deposits mainly with one-dimensional {111} orientation. At higher temperatures the rate of deposition was extremely slow and also led to the destruction of mica surface layers. The mirrors produced at about 200° c. gave two principal diffraction patterns, one when the beam was parallel to a basal axis of the mica cleavage plane, which repeated every 60° of azimuth, and a second at the azimuth 30° to each of these repetitions. These diffraction patterns show that the silver was deposited with {111} faces in contact with the mica cleavage face in two definite azimuth orientations, namely, with a [110] lattice row parallel to either of the two basal axes of the mica, one being obtained from the other by rotation through 180° about the (111) plane normal. This corresponds to crystals in four different orientations similar to the orientations described by Rüdiger (1937) for gold and palladium and by Andrashenko, and Tyapkina and Dunkov (1948) for silver on mica.

In Figure 1 (Plate I) the beam was in a direction parallel to the [$1\bar{1}0$] azimuth of one of the (111) Ag faces. A strong refractive index effect was present with most specimens. Indexing was carried out using the reciprocal lattice diagrams reproduced by Wilman (1940), but with the assumption that the spots on the $\frac{1}{6}$ in. diagonals were the product of double diffraction and submicroscopical twinning rather than due to the presence of a ' buffer ' lattice as suggested by Menzer (1938). Closely analogous patterns to silver on mica have been obtained from nickel oxide evaporated on to the basal plane of sapphire by Thirsk and Whitmore (1940), where a complete indexed pattern applicable to the present case is given for the [$1\bar{1}0$] azimuth.

Many of the specimens of silver on mica show a further two-dimensional orientation with a cube parallel to the main basal axis, and a cube face parallel to the mica cleavage face. The orientation was best observed on specimens with the mica substrate heated to 250°–350° c. The original of Figure 1 shows the [$\bar{1}00$] azimuth, (100) face diffractions.

Iron on Mica

Iron was evaporated on to mica substrate kept at temperatures between 150° and 400°, the iron filament being at bright red heat. The iron was cleaned initially by rubbing with 0000 emery and care was taken to free the iron from oxide by preheating the filament electrically in a vacuum. No two-dimensional orientation was observed, all the patterns being due to one-dimensional orientation, {110} and {111}, the {110} orientation predominating. All the evaporation experiments gave excellent mirrors, which subsequently exposure showed to be remarkably resistant to rusting.

Silver on the (111) *Face of Potassium Chloride*

For purposes of comparison with the behaviour of silver evaporated on pseudo-hexagonal muscovite cleavage faces silver was evaporated on to the ground and etched (111) face of potassium chloride. A two-dimensional orientation is known for the (100) face.

With the substrate at 200–300° c. a pattern similar to that of silver on mica was obtained, indicating the presence of the same hexagonal orientations.

Iron on the (111) *Face of Potassium Chloride*

Although Brück (1936) found it comparatively easy to produce two-dimensional orientations of iron on the cleavage face of rock salt only one-dimensional orientations were observed with iron on the (111) potassium chloride face and then only with the substrate at temperatures higher than 200° c. A strong {110} and some {200} one-dimensional orientation was present.

Platinum on to a (100) *Potassium Chloride Face*

A freshly cleaved specimen of potassium chloride was maintained at a temperature of approximately 250° c., and platinum evaporated from a filament at about white heat. Several hours were required for the deposition of a film of appreciable thickness. Patterns were taken with the beam along the [100] and $[1\bar{1}0]$ azimuth of the cube face. Diffractions due to the potassium chloride substrate could clearly be seen. The patterns showed that the platinum films were of a normal untwinned single crystal structure, the cube of the platinum edge being parallel to that of the cube edge of the substrate. The [100] azimuth pattern is shown in Figure 2 (Plate I).

Silver on Magnesium Oxide

Magnesium oxide offered interesting properties as a substrate since, although cubic and of the sodium chloride type, it is regarded as differing from a purely ionic crystal (Wells 1947). Silver was evaporated on to the (100) cleavage face maintained at temperatures of from 400° to dull-red heat. Below 400° the deposits were mainly random. The chief orientation was {100} with the cube edge parallel to the cube edges of the magnesium oxide. A representative pattern is shown in Figure 3 with the beam along the {100} azimuth of the magnesium oxide. Figure 3 is a pattern due to a {111} twinned silver lattice [100] azimuth. In addition there are present diffractions corresponding to the one-dimensional orientations {211} strong and {111} relatively weak. The {211} one-dimensional orientation has often been observed when face-centred cubic systems are evaporated on to comparatively inert substrates.

Nickel on Magnesium Oxide

Nickel was evaporated from a thin filament previously flashed *in vacuo* to free from oxide and then maintained at bright-red heat. No orientation of the films was found with the magnesium oxide substrates at temperatures below 500° c. and above that temperature the orientations observed were all of one degree. Figure 4 shows that the principal one-dimensional orientation are {110} strong, {200} fairly strong, {110} fairly weak, {111} weak.

Silver on Calcite

At temperatures above 400° c. two-dimensional orientations of silver on calcite were found similar to those reported by Rüdiger (1937) for Ag, Au and Pd with {110} and {111} faces in contact with the calcite cleavage faces (Figure 13 of his paper). Below 400° c. extremely complex patterns were obtained showing however the presence of fairly well developed crystals.

The results described have been summarized in the Table.

Metal	Substrate	Comment
Silver	Muscovite mica	Below 200° c. mainly {111} one-dimensional orientation. Above 200° c. principally four {111} two-dimensional orientations [1$\bar{1}$0] lattice row parallel to either of the axes in the mica basal plane. Between 250°–300° c. An additional well-marked {100} two-dimensional orientation with the cube edge of the silver parallel to the axis in the basal plane of the mica. Evidence of submicroscopical twinning. Andrashenko *et al.* (1948) report the same cube face orientation.
Fe	Muscovite mica	Between 150° and 400° c. {110}, {111} and {110} one-dimensional orientation only. Below 150° random patterns.
Ag	(111) KCl	Between 200°–300° c. Behaviour very similar to silver on mica at the same temperatures.
Fe	(111) KCl	No preferred orientation at temperatures less than 200° c. At 350° c. a strong {110} and some {200} one-dimensional orientation only. Compare with Fe on mica and contrast with the easily obtained two-dimensional orientation observed by Brück (1936) on the cleavage face of rock salt.
Ag	(100) MgO	Below 400° c. Deposits mainly random. 400° c. to red heat. Chief orientation cubic face and cube edge parallel to the cube face and cube edge of the magnesium oxide. Evidence of twinning even at the high temperatures. Strong {211} and weak {111} one-dimensional orientations.
Ni	(100) MgO	Below 500° c. Deposits random. Above 500° c. Orientation of one-dimensional {100} strong, {200} fairly strong, {110} fairly weak, {111} weak.
Silver	Calcite cleavage face	Below 400° c. Very complex patterns probably due to a number of fairly large crystallites in slightly differing orientations. Above 400° c. Patterns confirming Rüdiger's results (1937) but also showing evidence of twinning.

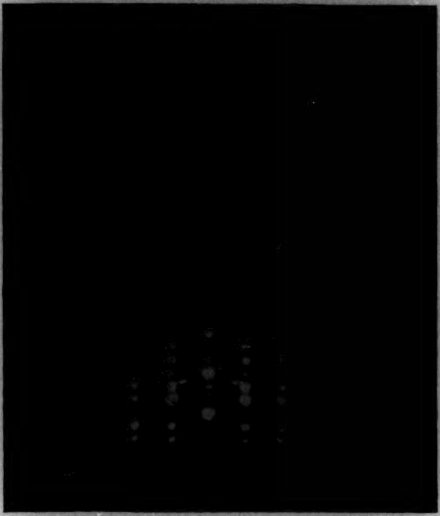


Figure 1. Silver on a mica cleavage face. Beam parallel to the [$1\bar{1}0$] azimuth of one of the (111) Ag faces and the basal axis of the mica.

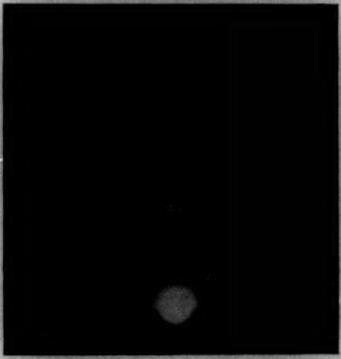


Figure 2. Platinum on a potassium chloride cleavage face. The beam parallel to the cube edge azimuth of the platinum and potassium chloride.

PLATE I.

To face page 836

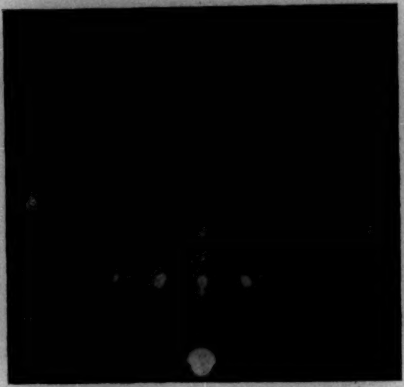


Figure 3. Silver on the cleavage face of magnesium oxide. Beam parallel to the cube edge of the principal silver crystals and the cube edge of the substrate crystal.

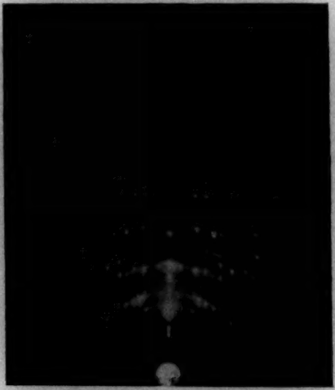


Figure 4. Nickel on the cleavage face of magnesium oxide.

PLATE II.

§ 3. DISCUSSION

The intention of this paper is primarily to describe some additional examples of the orientations assumed by metal films evaporated on to single crystal substrates, and these have been summarized in the Table; but in examining the results, difficulties of interpretation arise, the nature of which require some discussion. We may summarize them as follows: first, the extent to which the film is free from discontinuities, not merely in the sense of local dislocations, but in the likelihood that the deposit is formed of separate crystals of small size having uniform orientation; the extent to which the interfacial forces, assumed to be controlling the nature of the orientation, may have become a true chemical bonding; and finally, the question of how far one may consider that closeness of fit of the interatomic distances of the adjacent faces of metal and substrate deduced from the bulk orientation may have been decisive in the choice of the final atomic arrangement.

Electron diffraction cannot of itself provide a complete answer to the first problem. For a completely satisfactory examination recourse must be had to a supplementary technique using the electron microscope or, in certain problems, multiple beam interferometry. Hass (1942) has shown by electron microscopy that silver films grown on the (100) face of rock salt when less than 40 A. thick are in the form of isolated granules about 100 to 150 A. long. The diffraction patterns from the platinum films strongly suggest a similar structure. In addition a complete absence of any diffractions due to submicroscopical twinning is noted, although the slight arcing in patterns with the beam along the cube edge and cube face diagonal suggest that the crystals are extremely thin in a direction normal to the surface plane, and usually it is in this thin layer adjacent to the substrate that the strong evidence for twinning with face-centred cubic metals is found. There is a considerable difference between the a_0 spacing for platinum and potassium chloride: $a_0(\mathrm{Pt}) = 3{\cdot}916$ A., $a_0(\mathrm{KCl}) = 6{\cdot}278$ A.

Chemical interaction is in some cases an important factor in ensuring a surface orientation. Willems (see van der Merwe 1949 for reference) has quoted a number of examples, but a result particularly relevant to the work described here is an example studied by Thirsk and Whitmore (1940) concerning the orientation of Fe_2O_3 on a magnesium oxide substrate with the observed formation of an intermediate layer of spinel. It appears from a consideration of the behaviour of nickel on magnesium oxide and iron on mica, that chemical interaction may under certain conditions prevent rather than assist the production of a two-dimensional orientation. In these two cases chemical bonding may prevent further movement after condensation of the nickel and iron on the substrate. With analogous experiments with evaporated silver, the surface mobility is retained in the absence of chemical bonding.

The behaviour of magnesium oxide is particularly interesting owing to the diversity of its behaviour as an active substrate. It has been used as such by Vineyard (1942) for the ionic crystals LiF, NaF and NaCl deposited from solution and from the melt. There was an easily produced two-dimensional orientation only in the case of LiF, where the lattice fit is very close. In the present work the lattice fit of Ag on magnesium oxide is good ($a_0(\mathrm{Ag}) = 4{\cdot}077$ A., $a_0(\mathrm{MgO}) = 4{\cdot}203$ A.). We can summarize the behaviour of magnesium oxide as an active substrate. From Vineyard's work a controlling factor for ionic deposits is apparently a close lattice fit. From previous work by the author with a non-ionic deposit Fe_2O_3, chemical interaction and the formation of an intermediate compound can produce a two-dimensional orientation. The spinel interfacial compound shows signs of twinning. With nickel deposits reported in this paper chemical bonding may occur, two-dimensional orientation being prevented by the limitations thus imposed on the mobility of the nickel atoms in the first deposited layer. With silver a moderate degree of two-dimensional orientation is obtained at high temperatures due to Van der Waals surface forces but presumably without chemical bonding. There is evidence again of submicroscopical twinning.

Wilman (1940) has shown that silver films separated from the substrates and showing twinning spots may readily be annealed by heating to 500° c. Temperatures higher than this failed to produce annealing with the films of silver on magnesium oxide, suggesting that the surface forces at these higher temperatures must still be very considerable.

The patterns from silver on mica and on the (111) potassium chloride face are very similar to those described by Rüdiger (1937), but the extra diffractions are interpreted as being due to submicroscopical twinning and double diffraction.

Additional details of the lattice fits at the interface of the deposit and substrate implied by the orientation of the bulk deposit are as follows: for silver ($a_0 = 4{\cdot}06$ A.) having a $\{111\}$ face orientation on muscovite mica a percentage misfit of 10% must exist, using van der Merwe's nomenclature, if we assume that the silver ions fit into the hexagonal arrangement of spaces in the oxygen network as suggested by Andrashenko and co-workers (1948), or -3% if the position of the OH′ groups in the mica with a hexagonal spacing of 3·0 A. are of importance in deciding the orientation (for the structure of mica, see Jackson and West 1930). With silver on the (111) potasium chloride face ($a_0 = 6{\cdot}06$ A.) the misfit is of the order of 30%. Rhodin (1949) examining the orientation of aluminium on the (111) face of LiF and NaCl associates the poorer fit on the NaCl face with a greater disparity in the principal lattice dimensions. It was pointed out by Thomson (1948) in summarizing data on the orientations assumed by metal films that they are frequently not those giving the apparently most favourable lattice. Van der Merwe suggests that Menzer's geometrical picture (see Menzer 1936, 1938) of the establishment of a final orientation through the growing together of intermediate layers of different orientation having a good lattice fit at the substrate surface is a plausible treatment of the problem. This hypothesis explains the existence of certain extra spots in the diffraction patterns, but it has been shown that they may also be explained by less elaborate postulates concerning the nature of the deposit (Wilman 1940, Thirsk and Whitmore 1940). An alternative process would be through the primary formation of an approximation to a hexagonal close-packed layer followed by a change through adjacent layers to the final orientation, any strain produced being relieved by slip processes, for instance those giving rise with face-centred cubic metals to submicroscopical twins. With a process of this kind the nature of the symmetry of the rigid substrate face by defining positions of general lattice fit, having a symmetry similar to that of the substrate face, could cause the appearance of a similar symmetry in its final film. This is then maintained even though the agreement in spacing atom for atom along the main lattice rows at the interface is poor and liable to give rise to a large number of local dislocations in the films deposited.

ACKNOWLEDGMENTS

The author would like to record his thanks to Professor G. I. Finch for his interest and Dr. H. Wilman and Dr. E. J. Whitmore for much helpful discussion during the period when the experimental work described in this paper was being carried out.

REFERENCES

ANDRASHENKO, H. V., TYAPKINA, V. V., and DUNKOV, P. D., 1948, *Doklady Acad. Nauk, S.S.S.R.*, **59**, 1113.

BRÜCK, L., 1936, *Ann. Phys., Lpz.*, **26**, 233.

GOCHE, O., and WILMAN, H., 1939, *Proc. Phys. Soc.*, **51**, 625.

HASS, G., 1942, *Kolloid Z.*, **100**, 230.

JACKSON, W. W., and WEST, I., 1930, *Z. Kristallogr.*, **76**, 211.

VAN DER MERWE, J. H., 1949, *Discussions on Crystal Growth* (London : Faraday Society).

MENZER, G., 1936, *Naturwissenschaften*, **26**, 385; 1938, *Z. Kristallogr.*, **99**, 378, 410.

RHODIN, T. N., 1949, *Discussions on Crystal Growth* (London : Faraday Society), p. 215.

RÜDIGER, O., 1937, *Ann. Phys., Lpz.*, **30**, 505.

SHIRAI, S., 1937, *Proc. Phys.-Math. Soc., Japan*, **19**, 937; 1938, *Ibid.*, **20**, 855; 1939, *Ibid.*, **21**, 800; 1941, *Ibid.*, **23**, 12.

THIRSK, H. R., and WHITMORE, E. J., 1940, *Trans. Faraday Soc.*, **36**, 565, 862.

THOMSON, G. P., 1948, *Proc. Phys. Soc.* A, **61**, 403.

VINEYARD, S. M., 1942, *Phys. Rev.*, **61**, 100.

WELLS, A. F., 1947, *Structural Inorganic Chemistry* (Oxford : University Press).

WILMAN, H., 1940, *Proc. Phys. Soc.*, **52**, 323.

Surface Effects in Creep of Cadmium Crystals

BY D. J. PHILLIPS AND N. THOMPSON

H. H. Wills Physical Laboratory, University of Bristol

Communicated by N. F. Mott; *MS. received* 31*st May* 1950

ABSTRACT. A recording rate-of-strain meter has been used to investigate the changes in creep rate when the chemical environment of a stressed cadmium crystal is altered. Experiments were made with aqueous solutions of a number of inorganic salts of cadmium. The observations are explicable in terms of variations in the thickness of a surface film of cadmium hydroxide, the presence of such a film having the effect of reducing the creep rate by an amount depending on its thickness. Measurements were also made on the effect of very thin films, formed by immersion in distilled water; these were estimated to be about 10^{-6} to 10^{-7} cm. thick. When the films were removed with dilute sulphuric acid a sudden small strain increment was observed, the magnitude of which depended on the film thickness. Possible explanations are discussed.

§ 1. INTRODUCTION

THE work to be described was suggested by certain effects reported by Andrade and Randall (1948). In their work, clean single crystals of cadmium were made to extend uniformly at a rate of about $1{\cdot}7 \times 10^{-4}$ per minute, and were then surrounded by various inorganic salt solutions (mainly salts of cadmium). They found that the rate of strain suffered changes which were in some cases very large. For example, the application of cadmium sulphate solution increased the strain rate by a factor of 4; cadmium nitrate gave at first an increased rate, followed by a complete stoppage; with zinc nitrate, immediate hardening occurred.

Effects of surface conditions upon the strength of metal crystals have also been reported by Roscoe (1936), by Cottrell and Gibbons (1948) and by Harper and Cottrell (1950) who were concerned with oxide films. A considerable amount of work has also been done in Russia by the Rehbinder school. The fracture and cleavage strengths of mica and calcite, and the resistance to plastic deformation of metals (polycrystals and single crystals) were all found to be reduced when various surface-active substances (e.g. higher alcohols) were added to the medium surrounding the specimen (Rehbinder and co-workers 1931–1947). Some doubt however, exists as to the validity of some of this work (Kemsley 1949) and the question has been discussed by Harper and Cottrell (1950).

In order to study the variation in strength (i.e. resistance to plastic flow) of a crystal, two methods suggest themselves:

(*a*) The measurement of critical shear stress. A number of uniform crystals, of the same orientation, would be given various surface treatments and the stress required to produce a known, small strain rate measured.

(*b*) A given specimen could be made to extend under constant load, and the various surfaces prepared while the test was in progress. The effects of these changes on the rate of strain could then be studied.

The first method gives quantitative data: even with the best techniques, however, the measurement of changes of critical stress of 1% or less would require statistical experiments upon numerous samples. Moreover, the method is obviously not suited to continuous measurements. In some of the experiments to be described, surface conditions were not stable and the use of this method would have been impossible. The second method was chosen for the present work. It has the disadvantages that the crystal is subjected to strain-hardening during the experiment, and that it is difficult to obtain a quantitative measure of the effects. For the *detection* of the effects, however, the method is very sensitive.

In practical convenience the second method is much to be preferred. In one experiment many surface changes can be investigated (as many as 90 separate tests have been made). Further, since work-hardening must in any case be present, there is no reason why the same crystal should not be used for many experiments. In the present work, a crystal was not discarded until the limit of plastic flow was approached (about 40% extension). By its nature, the method is not affected by some factors which would be important in the first method. Changes in temperature or stress, for example, provided that they occur slowly, do not affect the results since only rapid changes in strain rate were measured. Nor was it necessary to take any great precautions regarding the initial alignment and straightness of the specimen, since fluctuations in creep rate due to these factors disappear after a small extension.

§2. EXPERIMENTAL DETAILS

The cadmium used was kindly supplied by the National Smelting Co. The impurities were as follows: lead 0·007%, zinc less than 0·01%, thallium 0·0005%, tin less than 0·0005%, copper less than 0·0001%, calcium trace: total less than 0·018%.

Crystals were grown in air by the method of Andrade and Roscoe (1937). They were of slightly oval cross section, of mean diameter 1·9 mm., and 60 cm.

long. They were cut into 10 cm. lengths for the experiments. The crystals, as grown, were somewhat oxidized. They were cleaned by electropolishing for ten seconds in 50% orthophosphoric acid, at a current density of about 1·5 amp/cm^2.

The results quoted in this paper were all obtained with portions of one crystal, in which the angle between the slip plane and the specimen axis was 25°, and the angle between the slip direction and the specimen axis was 36°. The crystal was supported vertically and loaded by hooks soldered to its ends, the loads being applied by means of a weight and lever. The strain was measured by recording photographically the movement of a spot of light reflected from a mirror fixed to the lever. The magnification was 16 times.

Since the work primarily required observation of the rate of strain, a device to record this directly was also used. A thread affixed to the lever passed over the pivot-shaft of a moving-coil milliammeter, and ended in a small weight. Movement of the lever caused the coil to rotate, and the resulting induced current actuated a galvanometer. The galvanometer deflection was therefore proportional to the velocity of the lever, and hence to the rate of strain of the specimen. This is true, of course, only if the strain rate is approximately constant. If the specimen elongates discontinuously in a time short compared with the galvanometer period, the system effectively integrates, and a transient deflection is obtained, proportional to the strain increment. A photographic record of the strain rate was made on the same camera as was used to record the strain. The velocity sensitivity of the arrangement used was 800 cm. deflection for 1 cm/sec. specimen velocity, the displacement sensitivity for sudden extensions was 6,300 cm. per cm. extension of the specimen.

The crystals were subjected to the action of the various reagents by directing a fine jet of solution on to the upper end of the specimen. Two jets were used, either of which could be directed upon the specimen at will by a slight rotation of their support. The rate of flow was 0·5 cm^3/sec. In the work on thin films (§ 4), where it was thought that the forces exerted by the jets on the crystal might vitiate the results, the jets impinged upon a small aluminium plate, and were conducted to the specimen by a projection of this plate which almost touched it. The liquids used were always within 0·5° c. of room temperature, and careful tests showed that errors due to thermal changes were not important.

It was found that any change in the composition of the solution surrounding the metal usually produced a change in the creep rate. The extent, and even the sign, of the change in rate depended not only on the nature of the new solution, but also on the previous history of the specimen. It appears that the changes in creep rate are to be correlated with changes in the thickness and/or composition of the film of hydroxide * which forms on the metal surface under these conditions. The experimental results fall conveniently into two groups distinguished by the order of magnitude of the thickness of the surface films involved. The results are presented separately in §§ 3 and 4.

* No attempts were made to determine the exact composition of the surface films formed. When they were thick enough to be visible they were white, and turned into the characteristic brown oxide on warming. Mellor (1929) speaks of films of a 'hydrated oxide' being formed on cadmium in water: for convenience the films are referred to as 'hydroxide' throughout. (Since the above was written Mentor and Hall (1950) have reported the occurrence of a crystalline overgrowth on oxide-coated cadmium crystals immersed in cadmium chloride or nitrate: from electron diffraction photographs this was identified as cadmium hydroxide.)

§ 3. RESULTS: THICK HYDROXIDE FILMS

To demonstrate the effects of relatively thick films of cadmium hydroxide, an electrolytic method of preparation was used. The specimens were anodically oxidized in an electrolyte of potassium hydroxide at a current density of about 1·4 amp/cm^2. The hydroxide film builds up in a few seconds to an equilibrium thickness of about 10^{-4} cm. (The film thicknesses quoted throughout were estimated by weighing, the density of the film being taken to be the same as that of bulk cadmium hydroxide. Because of this assumption the values of absolute thickness are liable to be in error, perhaps by a factor of two. The relative accuracy, however, is to within a few per cent.)

Figure 1, which is copied from one of the original photographic records, shows the effect upon the creep rate when such a film is removed. Before the point B the hydroxide-covered crystal was extending uniformly in distilled water at a low rate (about 3×10^{-5} per min.). At B, 2% sulphuric acid was directed upon the crystal. Cadmium metal is not rapidly attacked by acid of this concentration: a separate experiment showed that the rate of solution of the metal was about $4{\cdot}4 \times 10^{-6}$ cm. depth per minute. The hydroxide film, on the other hand, was removed within one second by conversion to the soluble sulphate. The rate of extension immediately increased by a factor of about 60; subsequently it slowly decreased again, due to the hardening of the crystal. At the point C, distilled water was substituted for the acid, with no sensible effect upon the strain rate. With more dilute acid (0·1%) the removal of the film and the consequent rise in strain rate took place more slowly, rising from 10^{-4} per minute to a maximum of $1{\cdot}5 \times 10^{-3}$ per min. in 1·3 minutes.

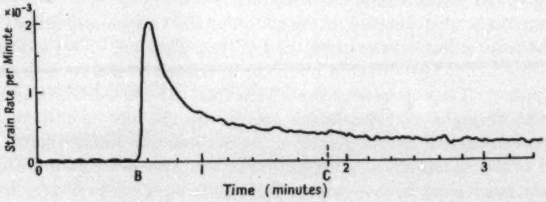


Figure 1.

Immersion of an electropolished crystal in a solution of cadmium nitrate was found to deposit hydroxide films and to reduce the creep rate. The removal of such films with sulphuric acid while creep was in progress produced effects entirely similar to those mentioned above. The films were easily visible; immersion for four minutes in a semi-normal solution produced a film 4×10^{-5} cm. thick. The effect of depositing a film upon a clean crystal while creep was in progress is shown in Figure 2. The crystal was extending in distilled water. At the point A semi-normal cadmium nitrate was applied, and the strain rate immediately began to decrease. At B 2% sulphuric acid was substituted for the nitrate. The removal of the film built up during the application of the cadmium nitrate solution caused the large increase in strain rate shown.

On the other hand if the crystal were already covered with a ‘ thick ’ hydroxide film, prepared by anodic oxidation as previously described, application of the same reagent (cadmium nitrate) caused an increase in the creep rate. This behaviour is shown in Figure 3. Curve ***b*** shows the effects upon the rate of strain when a semi-normal solution of cadmium nitrate was applied (at $t=0$) the crystal having been previously coated electrolytically with hydroxide. The strain rate rose by a factor of 6·3 in 2 minutes. Subsidiary weighing experiments showed that under these conditions the thickness of the film was reduced. At $t=147$ seconds, 2% sulphuric acid was applied. The effect is shown in curve ***a***. This has been drawn so that the instant of application again falls at the time origin. The rapid removal of the remainder of the hydroxide film causes a typical rapid increase in the strain rate followed by hardening.

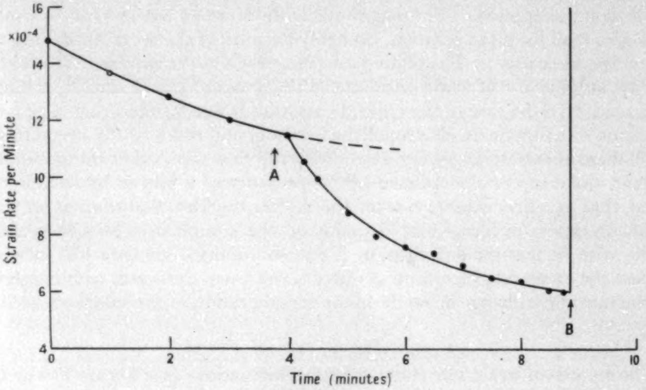


Figure 2.

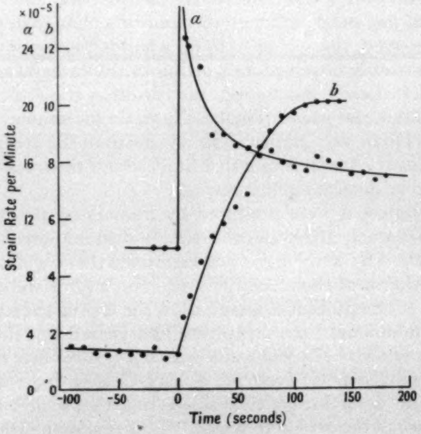


Figure 3.

Various other substances were found to behave in a similar manner, with partial removal of a pre-existing thick hydroxide film; the salts investigated were a commercial cadmium plating solution (Canning 'Zonax') and semi-normal solutions of cadmium chloride, nitrate and sulphate. The speed with which the film was removed varied in this order (from about 3 seconds for the first to 100 seconds for the last); the records obtained showed an increased creep rate when the solution was applied. The magnitude of the increase was less for slow solution of the film than for rapid solution, probably because of the work hardening taking place at the same time. The plating solution gave a curve similar to that obtained with 2% sulphuric acid while cadmium sulphate gave a curve similar to that with 0·1 % acid. In the case of the chloride solution it was verified that it caused the deposition of a film on an electropolished crystal, and reduced the creep rate.

All these observations are consistent with the view that, other things remaining constant, the creep rate is reduced by the presence of a film of hydroxide on the crystal, that the effect is more marked the thicker the film, that there is an 'equilibrium' thickness of film which depends on the composition of the solution in contact with it, that the substitution of one solution for another will increase or decrease the creep rate according as the film thickness decreases or increases, and that the rate of the change depends on the concentration of the solutions used.

§4. RESULTS: THIN FILMS

The records of strain rate show irregular fluctuations (see Figure 1) which may be due to friction in the measuring system, or may possibly represent real variations. In any case their presence means that small changes in strain rate cannot be detected with certainty. However, it was observed that when a thin film is removed rapidly from the surface of the metal, although the resulting change in creep rate may be too small to be detected, there is in addition a sudden extension at the instant of removal. This is readily observed and, owing to the integrating property of the strain-rate recorder already mentioned, the resulting transient deflection of the galvanometer is in this case proportional to the strain increment. The removal of such a film appears to be very similar in its effects upon the creep rate to a sudden small increase in load. It was found that films whose thickness was estimated to be only 20 A. could be detected in this way.

The films investigated were produced by immersing the crystal in distilled water. Cadmium immersed in water with access of air is known to form a hydrated oxide (Mellor 1929). By weighing, it was estimated that the film deposited in ten minutes had a thickness of about 3×10^{-6} cm. Such films were rapidly removed by sulphuric acid. The procedure was to allow the crystal to creep under constant load for a few minutes until the creep rate had ceased to fall rapidly. It was then exposed alternately to distilled water and to 2% sulphuric acid. A transient deflection of the galvanometer occurred at each change from water to acid, but none at the opposite interchange, thus suggesting that a film had been formed during the exposure of the crystal to water. This is shown in Figure 4, in which the changes from acid to water are marked by single ordinates, and those from water to acid by double ordinates.

If the changes of liquid were repeated many times at regular intervals—say once per minute—the magnitude of the transient showed a progressive fall as the crystal hardened. In order to allow for this, the applied load was momentarily increased at frequent intervals during the experiment. These load increments were, typically, 1% of the steady load and were applied for 3 seconds. They gave rise to a sudden strain increment, which appeared on the strain-rate record as a transient similar to that produced by removing the thin hydroxide film. By comparing the heights of the transients due to acid application with those due to a standard load increment, the effects of strain-hardening could be allowed for.

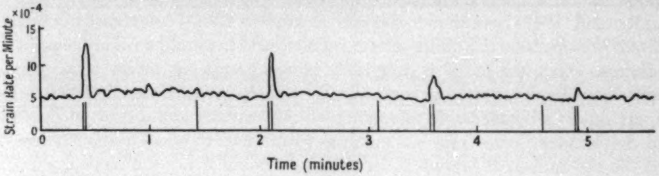


Figure 4.

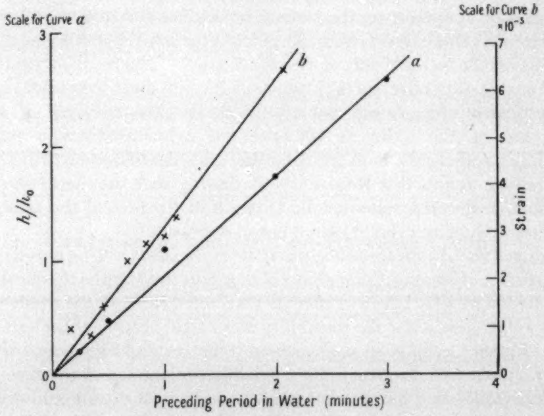


Figure 5.

In this way it was found that the magnitude of the sudden extension due to acid application increases when the duration of the previous immersion in water is increased. This dependence is shown in Figure 5 curve *a* in which ordinates are the ratio of the corresponding transient galvanometer deflection to the 'standard' deflection produced by a load increment of 0·2%. Figure 5 curve *b* shows a similar set of observations, obtained from a specimen which showed only very little strain-hardening during the time occupied by the measurements: the absolute magnitude of the transients can accordingly be given without ambiguity. The simplest interpretation of Figure 5 is that the film thickness increases linearly with time at least for the times shown. If this is true, then using the result already quoted for the estimated thickness of the film after 10 minutes immersion, we may conclude that the removal of a film about 10^{-7} cm. thick produced a measurable effect. This consisted of a sudden increase in strain of order of magnitude 10^{-5}; there may possibly have been also an increase in strain rate, but if so it was too small to be detected.

§ 5. DISCUSSION

In view of the results reported in §§ 3 and 4 it appears probable that the results of Andrade and Randall mentioned in the introduction would be explicable in terms of surface films formed or removed by the inorganic reagents which they used. It is also becoming clear that the conflicting results concerning the effects of organic liquids may be reconciled when due importance is attached to such films. Andrade and Randall (1949) have shown that the ‘ Rehbinder effect ’—increase in creep rate caused by immersion in a dilute solution of oleic acid in paraffin—is only found with cadmium when the metal is originally covered with an oxide film. Similar results have been reported in more detail by Harper and Cottrell (1950). Both papers report a considerable time interval between the application of the liquid and the increased creep rate—as much as 45 minutes in some circumstances—so that the mechanism of ‘ removal ’ of the film is possibly not the same as when the hydroxide is dissolved in strong acid.

In any event there remains the problem of exactly how the presence of a thin film of oxide or hydroxide on the surface can reduce the rate of creep to such a marked extent. One cannot reject out of hand the naive hypothesis that the film simply relieves the metal of part of the tensile load. The results already given in §§ 3 and 4 enable us to calculate the load which the film must have been supporting, and the estimates of thickness enable us to convert this to stress. The results obtained are of the order of 10^9 dynes/cm^2 which, although not entirely impossible, is sufficiently high to make the explanation somewhat improbable. It is interesting to note that Roscoe (1936), dealing with the effects of oxide films on cadmium, deduced a value for the strength of the film of the same order of magnitude although by a very different line of approach.

In the same way it appears unlikely that the explanation is to be found in thermochemical effects. Certainly the amounts of heat produced by the chemical reactions involved are too small to have any effect if distributed through the whole volume of the metal. But there is also the possibility that a local heating of the surface might suffice to ‘ trigger ’ a slip process which would then extend right across the crystal. Such for example would appear to be a possible explanation of the effect reported by Andrade (1945) that bombardment of the surface of a cadmium crystal with α-particles can cause a considerable increase in the creep rate. In the present experiments, however, approximate calculations show that the temperature changes involved are not sufficient to give rise to any such process.

A more promising suggestion is concerned with the influence of the surface film on sub-microscopic regions of stress-concentration near the surface of the metal. If we adopt current views, we are to describe processes of plastic deformation of metal crystals largely in terms of the movement of ‘ dislocations ’. It can be shown that the formation of new dislocations is an improbable process for ordinary values of applied stress, and is only likely to occur where local conditions provide a stress concentration. One place where such conditions might arise is on the free surface of a single crystal, and in so far as we must consider the formation of new dislocations at all, it is not unreasonable to suppose that the process might be considerably hindered by the presence of a solid film. Similarly, dislocations already existing within the crystal and approaching a free surface at low speeds might be held up by the presence of a film on the surface: the removal of the film would release all such dislocations at once and might account for the small sudden extensions (‘transients’) discussed in § 4. This suggestion receives some support from the fact that all the effects discussed in this paper are hardly noticeable with polycrystalline material, in which the importance of the free surface of the metal is eclipsed by the large number of grain boundaries which can also act as sites of possible stress concentration.

REFERENCES

Andrade, E. N. da C., 1945, *Nature, Lond.*, **156,** 113.

Andrade, E. N. da C., and Randall, R. F. Y., 1948, *Nature, Lond.*, **163,** 890; 1949, *Ibid.*, **164,** 1127.

Andrade, E. N. da C., and Roscoe, R., 1937, *Proc. Phys. Soc.*, **49,** 152.

Cottrell, A. H., and Gibbons, D. F., 1948, *Nature, Lond.*, **162,** 488.

Harper, S., and Cottrell, A. H., 1950, *Proc. Phys. Soc.* B, **62,** 331.

Kemsley, D. S., 1949, *Nature, Lond.*, **164,** 1127.

Mellor, J. W., 1929, *A Comprehensive Treatise on Inorganic and Theoretical Chemistry* (London: Longmans), **4,** 474.

Menter, J. W., and Hall, E. O., 1950, *Nature, Lond.*, **165,** 611.

Rehbinder, P., 1931, *Z. Phys.*, **72,** 191; 1947, *Nature, Lond.*, **159,** 866.

Rehbinder, P., Kalinowskaya, N., and Wenström, E., 1933, *Phys. Z. Sowjet*, **4,** 365.

Rehbinder, P., and Lichtman, V. I., 1941, *C. R. Acad. Sci. U.R.S.S.*, **32,** 130.

Rehbinder, P., Lichtman, V. I., and Maslenikov, V. M., 1941, *C. R. Acad. Sci. U.R.S.S.*, **32,** 125.

Rehbinder, P., and Logginow, G., 1941, *C. R. Acad. Sci. U.R.S.S.*, **30,** 491.

Rehbinder, P., and Wenström, E., 1937, *Bull. Acad. Sci. U.R.S.S.*, **4,** 531; 1944, *Acta Physicochimo*, **19,** 36.

Roscoe, R., 1936, *Phil. Mag.*, ser. 7, **21,** 399.

Creep in a Precipitation-Hardened Alloy

By M. DAVIS and N. THOMPSON
H. H. Wills Physical Laboratory, University of Bristol

Communicated by N. F. Mott; *MS. received* 15*th March* 1950

ABSTRACT. Measurements have been made of the creep shown by polycrystalline wires of a hardened alloy (Cu+3% Ag) both at room temperature and 90° K. The results are interpreted in terms of the Mott–Nabarro theory of transient creep. The agreement between theory and experiment is qualitatively good, and the quantitative discrepancies suggest directions in which the existing theory might be refined.

§1. INTRODUCTION

Mott and Nabarro (1948) have published a theory of transient creep in metals based on the idea of the gradual exhaustion of a supply of dislocations already present in the metal crystal at the beginning of the test. The quantitative conclusions of the theory are, for reasons more fully explained in §2, applicable in their simple form to a ‘ hard ’ alloy under small stresses. Since no data were available in a convenient form for comparison with the theory, the experiments described in this paper were undertaken to provide such data; the results obtained have accordingly been interpreted throughout in terms of the concepts of this theory.

For convenience we give in §2 a brief resumé of the theory. The mode of presentation is not that of the original paper but seems better adapted to the interpretation of the subsequent experimental results.* A brief account of some of the experimental details is given in §3 and the results obtained are presented in §4.

§2. THEORY

A fundamental assumption is that any real metal crystal contains a large number of dislocations present in the lattice and that the movement of these, under the action of applied stresses, constitutes the plastic deformation. Following Taylor (1934) it is further assumed that the motion of any dislocation, or loop of a dislocation, is impeded by a complex system of internal stresses arising from impurity atoms, precipitates, other dislocations, grain and mosaic boundaries, etc. Considering any one dislocation loop, one supposes that it will move only when the force on it, due to the externally applied stress, becomes greater than that due to the system of internal stresses, i.e. each loop will have a characteristic stress σ which is just sufficient to set it in motion. Consequently the system of internal stresses may be characterized by a distribution function, $N(\sigma)$ such that $N(\sigma)d\sigma$ specifies the number of dislocations per unit volume which can only just be set in motion by an applied stress between σ and $(\sigma + d\sigma)$. The variation of the function $N(\sigma)$ with σ is, of course, not known, but may well be of the form sketched in Figure 1.

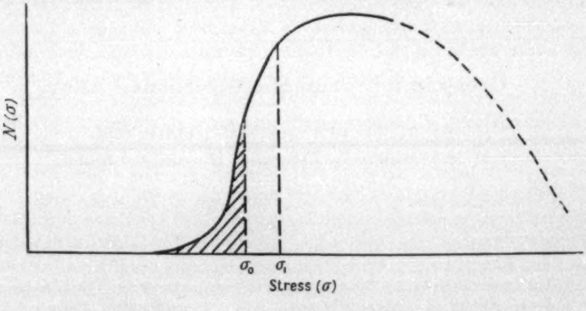


Figure 1. $N(\sigma)\,d\sigma$ is the number of dislocation loops per unit volume which would just move, at absolute zero, under an applied stress σ.

The application of a certain stress σ_0 will cause the immediate movement of all dislocations of characteristic stresses $\sigma \leqslant \sigma_0$ (i.e. those in the shaded portion of Figure 1). The immediate movement of these dislocations constitutes the 'instantaneous' plastic deformation. Dislocations which require an applied stress somewhat larger than σ_0 may possibly move in the course of time with the aid of thermal energy fluctuations. The plastic deformation so caused will constitute transient creep. Making certain plausible assumptions, Mott and Nabarro obtain the following expression for the activation energy for the motion of a loop of a dislocation, of length $\frac{1}{2}\Lambda$:

$$U(\sigma) = 0{\cdot}15\sigma a\Lambda^2(1-\sigma_0/\sigma)^{3/2}, \qquad \ldots\ldots(1)$$

---

* The method of treatment is due essentially to Professor Cottrell and was presented by him to a colloquium in Birmingham in 1948; it is published here for the first time with his kind permission.

where a is the lattice parameter, σ_0 is the applied stress, and σ is the applied stress under which the loop would just move immediately.

Dislocations will not be stationary in the lattice, but will vibrate about their equilibrium positions with a frequency ν, which is estimated theoretically to be of the order of 10^8 or 10^9 per second. (Much of the discussion of the results of this investigation is concerned with the value of this quantity.) Thus the chance $\alpha\, dt$ that the dislocation loop moves forward in a time dt is given by

$$\alpha = \nu \exp\{-U(\sigma)/\boldsymbol{k}T\}. \qquad \ldots\ldots(2)$$

Eliminating $U(\sigma)$ between (1) and (2), we have

$$\alpha = \nu \exp\left\{-\frac{B}{\boldsymbol{k}T}\left(1-\frac{\sigma_0}{\sigma}\right)^{3/2}\right\}, \qquad \ldots\ldots(3)$$

where B is written for $0{\cdot}15\sigma a\Lambda^2$.

It follows from the discussion in the original paper that the chance of a dislocation moving backwards (i.e. against the applied stress) is negligibly small. It is further assumed that once a dislocation has moved forward in this manner, the value of U appropriate to its new position will usually be too large for further motion to be possible, so that it has effectively been used up. Thus the number of dislocations in a stress range σ to $(\sigma + d\sigma)$ will decay according to the law

$$N(\sigma,t) = N(\sigma,0)\exp(-\alpha t).$$

At any instant t, therefore, the distribution function for the dislocations remaining can be calculated from the original distribution function by multiplying it by $\exp(-\alpha t)$, where α depends on σ according to equation (3). Insertion of reasonable numerical values shows that the function $\exp(-\alpha t)$ changes from a very small value to a value of almost unity over a short range of σ. The point of inflection on this step-like function is closely given by $\alpha t = 1$. As time proceeds, the position of this point of inflection changes in the direction of increasing σ. Without introducing serious errors, the function may be replaced by a sudden step, which advances along the axis of σ. The rate of advance is given by setting $\alpha = 1/t$ in equation (3). If σ_1 denotes the stress to which the step has advanced in a time t, we get, on rearrangement

$$\left\{\frac{\boldsymbol{k}T\ln \nu t_1}{B}\right\}^{2/3} = \frac{\sigma_1-\sigma_0}{\sigma_1} \simeq \frac{\sigma_1-\sigma_0}{\sigma_0} \qquad \ldots\ldots(4)$$

if σ_1 is not very different from σ_0.

The creep strain will depend on the number of dislocations which have moved, and if we assume that the motion of each one gives rise to an average strain v/V where V is the volume of the crystal, then the total creep strain is

$$S = v\int_{\sigma_0}^{\sigma_1} N(\sigma,0)\,d\sigma.$$

If we further assume that, over the limited range of σ concerned, we may take the original distribution function $N(\sigma, 0)$ to be independent of σ and equal to N_0, then

$$\begin{aligned} S &= vN_0(\sigma_1-\sigma_0) \\ &= vN_0\sigma_0(\boldsymbol{k}T\ln \nu t/B)^{2/3}. \end{aligned} \qquad \ldots\ldots(5)$$

Equation (5) is the same as that derived by Mott and Nabarro for the same conditions. With rather greater generality the distribution function may be taken as

$$N = N_0(1 - \sigma_0/\sigma)^m; \qquad \ldots\ldots(6)$$

the resulting creep equation is then essentially the same as (5) except that the index $2/3$ is replaced by $2(m+1)/3$.

One feature which is essential to the validity of this argument is that the distribution $N(\sigma)$ is unchanged throughout the experiment, except by the exhaustion process under consideration. The theory will not therefore be directly applicable to any material which shows appreciable re-softening or self-annealing at the temperature of the creep test. This follows since any such process can be thought of as an increase in the number of dislocations which will move under the lower stresses.

Similarly we must also exclude, for the present, materials in which an important part of the system of internal stresses is due to the presence of the dislocations themselves; in this case the movement of dislocations in the lower ranges of stress will considerably alter the forces acting on those remaining, and the distribution function will change continually as creep proceeds. This difficulty can be minimized if we set up a more important system of internal stresses, which outweigh the effects of the varying stresses due to work hardening and are themselves unaltered by the slip process, at least for small strains. It was thought that this requirement could best be met by an alloy hardened by internal oxidation or by incipient precipitation. In addition, the observations should, for preference, be confined to comparatively small strains.

For these reasons it was decided to carry out careful creep measurements on a precipitation hardened alloy. The system copper–silver was chosen as being suitable, since the precipitation process has been fairly well investigated and is not too complex. It is suppressed completely at room temperature, so that observations can be made on a stable material. Moreover, the comparatively high melting point suggests that appreciable self-annealing is unlikely to take place at room temperature.

§ 3. EXPERIMENTAL

All the results given in this paper refer to polycrystalline wires of a copper–silver alloy containing 2·75 per cent by weight of silver. The copper–silver alloy was prepared from high purity, spectroscopically standardized materials and was supplied by Messrs. Johnson, Matthey in the form of hard drawn wire, approximately 1·5 mm. in diameter. Specimens 12 cm. in length were annealed for three days at the eutectic temperature (780° c.) to ensure complete solution of the silver constituent. Both this solution and subsequent ageing heat treatments were carried out in an atmosphere of hydrogen in a tubular electric furnace. Each operation was terminated by water quenching. After surveying available x-ray data, a standard ageing heat treatment of 30 minutes at 500° c. was chosen to ensure that continuous precipitation with no crystal fragmentation would occur. The polycrystalline specimens so prepared had a mean grain size of 0·1 mm.

Attempts to grow single crystals of the alloy by the Bridgman technique were unsuccessful; the crystals were rarely single and there was always a pronounced segregation of silver to one end. Some degree of success was achieved by growing copper single crystals, electroplating them with appropriate weights of silver, and then subjecting them to a prolonged anneal at a high temperature. Alloy single crystals produced in this manner always had a slightly non-uniform silver concentration along their radius and were accordingly not considered suitable for quantitative creep measurements.* Some observations were made on them, however, and these reproduced all the characteristic features of the behaviour of the polycrystalline wires upon which most of the measurements were made. This provided confirmation for the belief that the presence of the grain boundaries would have a negligible effect on the general behaviour of the material at the temperatures concerned, which were all low compared with the melting point.

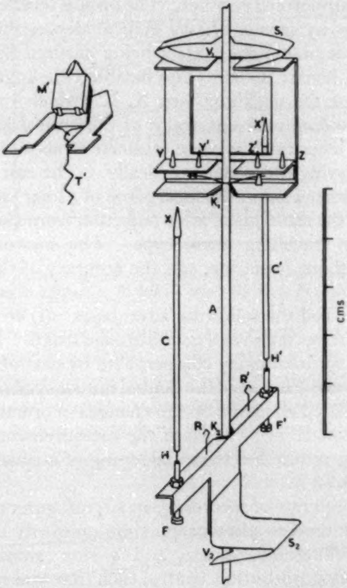


Figure 2. Extensometer; both mirrors removed and one shown inset.

The specimens were extended in a simple tensile machine, designed to apply loads up to 20 kg. As investigations were to be confined to small strains ($<1\%$) no provision was made for continuous compensation of load to maintain constancy of stress. The lower end of the specimen was held in a universal joint, and the stress applied to the upper end by a flexible wire, every care being taken to ensure conditions of axial loading. The apparatus was designed to enable the specimen to be surrounded by a copper can which in turn could be immersed in a constant temperature bath.

The light-weight extensometer (11 gm.) developed for this work was of the optical lever type, and measured the extension of a 5 cm. gauge length. Figure 2 shows a drawing of the extensometer mounted on a specimen, and with the two optical levers removed for clarity: one of them is sketched alongside. The gauge length was defined by the knife-edge V's (K_1 and K_2) each carried by a clip held firmly to the specimen by a light steel spring: one such is shown just above K_1. The steadying V's (V_1 and V_2) rested only lightly on the specimen, being cut in thin sheet steel springs S_1 and S_2 which needed only a very small force to move them parallel to the specimen axis. The lower clip carried two vertical comparator rods, C, C′, attached to two adjustment screws, F, F′, by short lengths of steel wire, H, H′. The four needle points on the upper clip were in a line intersecting the specimen axis. The needle points, X, Y, Z, supported one of the optical levers; the remaining three supported the other. The optical levers were located uniquely on the needle points by means of two conical depressions and one V-groove punched in each lever base, firm contact being ensured by light springs T, T′, which coupled on to hooks, R, R′. The flexible couplings H, H′, provided the necessary freedom for the needle points, X, X′ (which were located in conical depressions) thus avoiding over-constraint of the optical levers.

---

* The small scale chemical analysis necessary to reveal this was kindly undertaken for us by Messrs. Thos. Bolton and Sons, to whom our thanks are due for this service.

The two optical lever mirrors were inclined at about 45° to the horizontal, light entering and leaving the system vertically. The rest of the optical system consisted of an illuminated slit in the focal plane of a long focus lens ($f=120$ cm.); the image, formed in the same plane after reflection from the two mirrors in turn, was observed with a travelling microscope. The instrument could measure strains up to 10^{-2} without re-setting, and the accuracy of the values of strain was about 10^{-6}.

This arrangement had the following advantages : (i) errors due to end effects were eliminated by measuring over a central gauge length; (ii) thermal expansion of the specimen was approximately compensated by that of the comparator rods, made of similar material; (iii) use of the double mirror system not only doubled the sensitivity, but eliminated all errors due to changes in orientation of the specimen relative to the tensile machine; (iv) since the measurements were virtually made on the specimen axis, errors due to straightening of a specimen initially slightly bent were minimized.

In order to get a high rate of recording, an aircraft gun camera, adapted to take single exposures, was used to photograph simultaneously coarse and fine scales coupled to the travelling microscope, and a stop watch. The camera was electrically operated by push-button control each time the microscope cross-wire and the image were made to coincide. Between 20 and 30 accurate readings per minute could be obtained with this arrangement.

In some cases of transient creep, the initial creep rate was too great to enable the travelling microscope to follow the image and the first reading was only possible about a minute after commencement of creep. To obtain a record of this rapid creep, when the image was displaced several centimetres, a length of sensitive 35 mm. film was placed to receive the moving image. A rotating disc with cut-out sectors, driven by a synchronous motor, was placed in front of the illuminated slit. The interruptions served as time markers, so that 1, 1/4, or 1/10 of a revolution could be distinguished. The period of rotation was 1·1 seconds: thus intervals as short as 0·11 second were recorded. When the image displacement became slow enough for microscope observation, the film holder could be swung out of position and microscope settings made in the ordinary way.

§ 4. RESULTS

(*a*) *Creep Tests on Virgin Specimens*

Preliminary observations of strain made with a slowly increasing stress applied to a specimen showed that the material had a fairly well defined yield point. When the stress was very nearly equal to the yield stress, the creep behaviour was complex; the creep rate after falling from its initial value rose to a maximum before finally decaying to zero. This behaviour was only shown over a narrow range of applied stress. Further experimental work to elucidate the mechanism is in progress and it will not be considered further here.

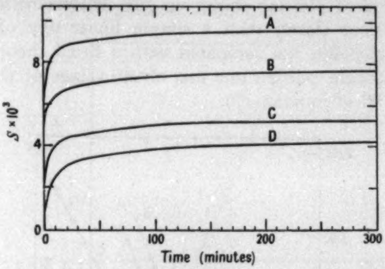


Figure 3. Creep of polycrystalline wires at room temperature. Stresses in kg/mm²: A, 6·55; B, 6·00; C, 5·77; D, 5·60.

At rather higher stresses the behaviour was normal. Figure 3 shows a number of graphs of strain plotted against time for polycrystalline specimens at room temperature. Each curve refers to a wire which had not previously been subject to any plastic deformation and which was loaded to the stress shown in the figure. Only the early parts of the curves are plotted, although the observations were usually continued for at least 24 hours. The creep rate decreased continuously during this period. Figure 4 shows the early portion of another similar creep curve on an

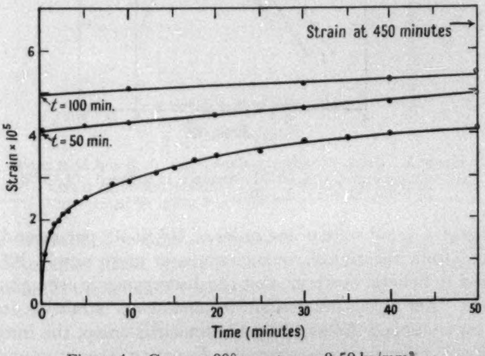


Figure 4. Creep at 90° A.; stress = 9·50 kg/mm².

enlarged scale; this illustrates the accuracy with which the observed points lie on a smooth curve. Many more observations were in fact available than are plotted even on this graph.

The curves of Figure 3 are of the typical transient creep type.* Similarly shaped curves were obtained for wires extended at liquid oxygen temperature, although the stress level was about twice as great. In order to see if the observations could be fitted to an equation similar to (5), graphs were plotted of strain to the power 3/2 against the logarithm of time; the result should be a straight line. Figure 5 shows a typical set of curves. These are straight for large times ($t > 10$ minutes) but show a curvature in the early stages, being concave towards the axis of time. This method of plotting shows up any disagreement between experiment and theory more clearly than a simple linear plot of (strain, time) in which experimental points are compared with a fitted theoretical curve. By extrapolating the straight portion one can obtain values of the constants ν and $K\{=vN_0\sigma_0(\boldsymbol{k}T/B)^{2/3}\}$ of equation (5).

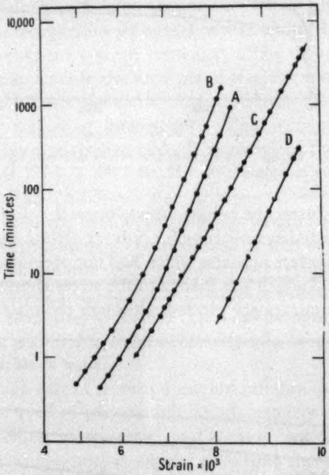


Figure 5. Creep of polycrystalline wires, A, B and D at room temperatures, C at 90° A. Stresses in kg/mm²: A, 5·70; B, 6·00; D, 6·55; C, 10·2.

The values of ν found were of the order of 10^2 to 10^4 per second, whereas the value expected from theoretical considerations is more nearly 10^9 per second. There is reason to believe, however, that the discrepancy is even greater than this would indicate. The theory predicts an 'instantaneous' strain occurring immediately the stress is applied, followed by the transient creep, the initial creep rate being very large but decreasing very rapidly. Equation (5) only attempts to account for the transient component of this creep. Now, in practice it is impossible to determine where instantaneous strain ends and transient creep begins, even with the photographic film method of recording mentioned in § 3, because a finite time is required to apply the stress, and the extensometer has a finite response time.

---

* All strains, as plotted, are plastic components; elastic components have been subtracted.

Consequently, the strains as measured should be reduced by an unknown amount to allow for the instantaneous strain, before the experimental data are compared with theory. The effect of this is two-fold: firstly, the curvature of the graphs of the type shown in Figure 5 is made somewhat more pronounced*, and secondly, the values of ν obtained by extrapolating the linear portion are reduced even further. It is noteworthy in this connection that those experiments giving the highest values for ν were those where the stress was highest and therefore the error due to the inclusion of the instantaneous strain was also greatest.

This difficulty is sometimes avoided by treating the instantaneous strain as an extra disposable constant in fitting an equation to the experimental data (cf. Andrade 1910, 1914). In the present instance, in view of the good accuracy of the observations of strain, it seemed preferable to analyse the results in terms of the strain-rate rather than strain, whereupon the difficulty disappears. By differentiating equation (5) one obtains immediately,

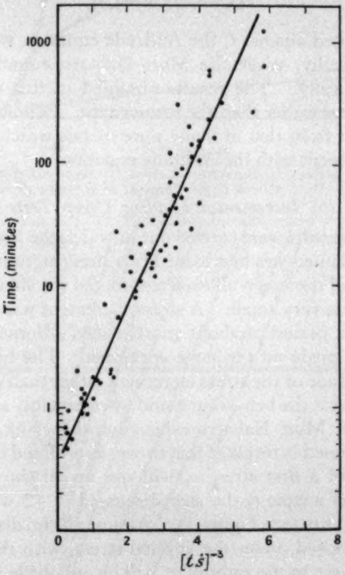


Figure 6. Creep at room temperature; stress $=6{\cdot}00$ kg/mm². Strain in arbitrary units. The data are the same as those represented by curve B of Figure 5.

$$(t\dot{S})^{-3}=\text{const.}\,(\log\nu+\log t),$$

which suggests plotting $(t\dot{S})^{-3}$ against $\log t$. The errors involved in determining the creep rate are, of course, magnified when $(t\dot{S})$ is raised to the third power, and thus the points show a considerable scatter. Figure 6 shows a typical curve.

---

[*] The linearity can be improved by plotting the strain to some higher power than 3/2, but this cannot reasonably be done within the framework of the Mott–Nabarro theory, since it would require $m<0$ in equation (6) and hence $N=\infty$ when $\sigma=\sigma_0$.

Estimating a ' best ' straight line through these points gives a value for ν of about 1 per second.

It is interesting to note in passing, that this method of analysing the results gives a sensitive method of distinguishing between the Mott–Nabarro equation and the Andrade equation (Andrade 1910, 1914). Under the conditions of these measurements, Anrade's ' *k*-flow ' is likely to be negligibly small, and his equation thus reduces to :

$$L = L_0(1+\beta t^{1/3}),$$

i.e.

$$S = \beta t^{1/3} \text{ for small strains.}$$

On differentiating this gives :

$$(t\dot{S})^{-3} = (3/\beta)^3(1/t).$$

Thus if $(t\dot{S})^{-3}$ is plotted against t, the Andrade equation would predict a curve decreasing hyperbolically, while the Mott–Nabarro equation predicts a curve increasing logarithmically. The results obtained in this investigation are unmistakably of this latter rather than the former type. The behaviour of this alloy is thus quite different from that of many pure metals which have given transient creep curves in agreement with the Andrade equation.

(*b*) *Incremental Loading Creep Tests*

A number of creep tests were carried out in which the loading programme was as follows. The specimen was first loaded to a stress at, or just above, the macroscopic yield point, and the creep allowed to proceed for about 24 hours. By this time the creep rate was very small. A stress increment was then made and creep readings taken over a period of about another day. Sometimes two successive increments could be made on the same specimen. The behaviour appeared to depend on the magnitude of the stress increment rather than on the absolute stress level. It was found that the behaviour could be reasonably accounted for in terms of the concepts of the Mott–Nabarro exhaustion theory of creep and the results obtained will be discussed in terms of that theory as outlined in § 2.

The application of a first stress σ_0 will use up all the most easily movable dislocations and, after a time t_1 the step discussed in § 2 will advance along the axis of σ to a point σ_1, say (see Figure 1). Almost all the dislocations lying to the left of σ_1 will have moved under the applied stress, with the aid of temperature fluctuations, while those to the right of σ_1 will be only little affected. The actual advance $(\sigma_1-\sigma_0)$ will depend on the nature of the material, the temperature, and the time. It has already been shown that the advance is approximately proportional to $(\log \nu t)^{2/3}$, so that after 24 hours (as in this experiment) it will be moving only slowly.

The material might now be expected to show fairly marked yielding for increments of stress greater than $(\sigma_1-\sigma_0)$ with an ' instantaneous ' plastic strain—as in the case of a virgin specimen stressed above its yield point—followed by a rapid, but rapidly decreasing creep. If the stress increment is less than $(\sigma_1-\sigma_0)$, however, there should be no instantaneous plastic strain, and the initial creep rate should be much lower.

It was to obtain evidence upon this point that the method of photographic recording, described at the end of § 2, was devised. The stress increments were applied as quickly as possible without causing any jerks, but, even so, observations in the first half-second are probably not reliable. The results obtained for three different stress increments are shown in Figure 7. It will be seen that two of them ($\Delta\sigma = 1{\cdot}60$ and $\Delta\sigma = 1{\cdot}86\,\mathrm{kg/mm^2}$) give strain curves showing a very rapid initial strain followed by a typical transient creep curve, whereas the other ($\Delta\sigma = 1{\cdot}20\,\mathrm{kg/mm^2}$) has quite a small initial creep rate. From a series of such observations, it was concluded that the 'yield point' for such incremental loadings was at about $1{\cdot}4\,\mathrm{kg/mm^2}$ when the previous stress had been applied for about 24 hours.

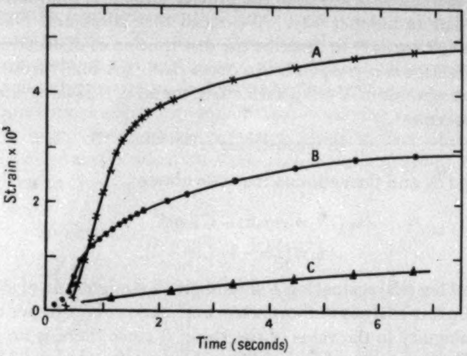


Figure 7. Initial parts of creep curves for incremental loading, room temperature. Stress increments in $\mathrm{kg/mm^2}$: A, 1·86; B, 1·60; C, 1·20.

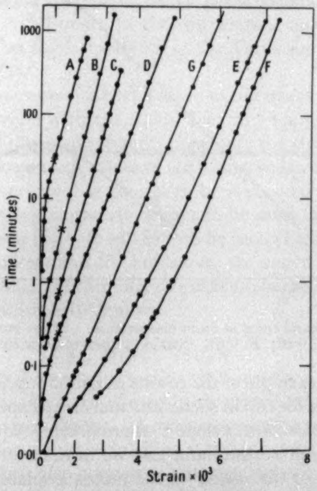


Figure 8. Creep curves for incremental loading, room temperature. Stress increments in $\mathrm{kg/mm^2}$: A, 0·10; B, 0·45; C, 1·01; D, 1·20; E, 1·60; F, 1·80; G, 1·40.

The results of such a series of tests are shown in Figure 8, in which strain to the power 3/2 is plotted against log (time) as before. For stress increments greater than the critical value the shape of the curves is the same as that obtained with a virgin specimen—concave towards the time axis initially and tending to become linear for large times. For increments less than the critical value the curvature has the opposite sign, while when $\Delta\sigma = 1{\cdot}40$ kg/mm² the points lie on a straight line.

When the increment is less than the critical value it is instructive to plot the experimental data in another way. We recall that Mott and Nabarro adopt a function, $N = N_0(1 - \sigma_0/\sigma)^m$ to describe the distribution of dislocations in terms of the stress σ. A large m corresponds to a curve rising gradually from zero at $\sigma = \sigma_0$, while $m = 0$ corresponds to $N = N_0$, a constant, for all σ. Making this assumption, equation (5) becomes

$$S = \text{const.}\,(\log \nu t)^{2(m+1)/3}. \qquad \ldots\ldots(7)$$

Differentiating (7), and then eliminating ν we obtain

$$\log(t\dot{S}) = \text{const.} + C \log S, \qquad \ldots\ldots(8)$$

where

$$C = (2m-1)/(2m+2).$$

Thus a graph of $\log(t\dot{S})$ against $\log S$ should give a straight line of slope C. Note that in applying this analysis to the incremental creep curves under consideration, there is no ambiguity in the value of the strain S since there is no instantaneous strain. The limiting values of C are: $C = -\frac{1}{2}$ for $m = 0$ and $C = +1$ for $m \to \infty$.

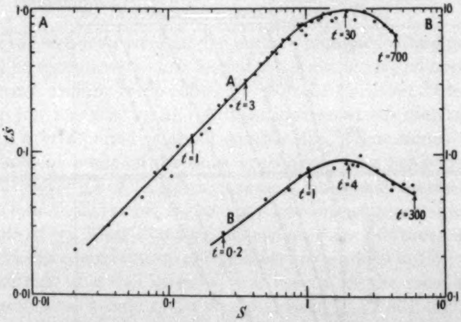


Figure 9. Incremental creep at room temperature. Stress increments in kg/mm²: A, 0·10; B, 0·64. Strain in units of $5{\cdot}4 \times 10^{-4}$.

Figure 9 shows an example of the results obtained when $(t\dot{S})$ is plotted against S on a logarithmic scale for two in suchcremental creep experiments. In the early stages the graphs are linear with a slope C approximately equal to 1; subsequently the curve passes through a maximum, and we have eventually a slope given by $-1 < C < 0$. In view of the scatter of the points available, it is not possible to determine the value of the negative slope at all accurately, but the data are not inconsistent with its being $-\frac{1}{2}$.

The immediate conclusion is that no simple law, such as equation (6), can describe the distribution function $N(\sigma)$ as it existed before the stress increment was applied. We would indeed expect it to be largely determined by the form of the step function $\exp(-\alpha t)$ corresponding to the initial stress. The problem of calculating the creep curve with a distribution function of this type appears to be quite intractable. One point may, however, be made. The treatment of exhaustion creep theory given in §2 shows that the creep rate at any time is determined almost entirely by the shape of the distribution curve near the value of σ to which the step has advanced, provided that $N(\sigma)$ is not varying rapidly with σ at this point. In the incremental loading tests under discussion, the step will eventually advance beyond the escarpment in the distribution function into the region where $N(\sigma)$ will not vary rapidly. The time taken to reach this situation will be less when the increment is greater, and will be zero when the increment equals to $(\sigma_1 - \sigma_0)$. It is suggested that the maximum in the graph of log $(t\dot{S})$ against log S corresponds to the transition from a regime in which $N(\sigma)$ is varying rapidly, to a regime in which $N(\sigma)$ is much more nearly constant. It is in fact observed that this transition takes place earlier when the stress increment is greater. This is shown by the data given in the Table.

Specimen	$\Delta\sigma$ (kg/mm^2)	t for $(t\dot{S})_{\max}$ (minutes)	S at $(t\dot{S})_{\max}$ ($\times 10^3$)
P 23 F	0·10	30	1·05
P 24 C	0·37	25	1·0
P 24 B	0·45	7	1·2
P 23 C	0·64	4	1·15
P 31 B	1·01	½ to 1	1·2

One would also expect that the total creep strain up to this point would be the same whatever the stress increment, since it corresponds to using up all the dislocations up to the escarpment on the distribution curve. This is fairly well borne out in practice, as the last column of the Table shows the strain at $(t\dot{S})_{\max}$ to be reasonably constant.

Only when this maximum has been passed would one expect the Mott–Nabarro creep equation to be valid with the index 2/3; in Figure 8 we observe that the relevant curves (A to D) become straight lines only for strains greater than those corresponding to the maximum values of $(t\dot{S})$—marked in Figure 8 by crosses.

When the stress increment is chosen to be just equal to the critical value of 1·4 kg/mm. one might expect the conditions to be most favourable for obtaining agreement with equation (5). It will in fact be seen (Figure 8, curve G) that the linear relation between log t and $S^{3/2}$ holds over the entire range from 10^{-2} to 10^3 minutes. By extrapolation of this line, one can obtain the value of ν more reliably than heretofore; the result is 10^2/minute.

§ 5. DISCUSSION

It will be clear from the previous section that the effects observed in the experiments on incremental loading are all readily accounted for in a qualitative manner in terms of exhaustion creep theory. A more quantitative treatment is difficult and it is doubtful if the effort would be justified at this juncture. The general form of the strain–time relation is also consistent with the theory.

When we consider the quantitative results however, the position is less satisfactory. The surprisingly low value of ν—the frequency of vibration of a dislocation loop—has already been remarked. If however, we accept this result, then a knowledge of the distance that the ' step ' advances along the axis of stress in a given time enables us, with the aid of equation (4), to evaluate the quantity B of equation (3). Taking the advance $(\sigma_1 - \sigma_0)$ to be 1·4 kg/mm^2 in 24 hours, one finds that B is of the order of $40\boldsymbol{k}T$—a not unreasonable figure. Further, from equation (1) setting $\sigma = 8$ km/gm^2—the order of magnitude of the stresses used—we find that Λ, the length of a dislocation loop, is about 60 A.: this too is not impossible, although smaller than might have been expected. Finally, the slope of a graph of $S^{3/2}$ against $\ln(\nu t)$ then enables us (equation (5)) to obtain a numerical value for vN_0. The incremental loading curve for $\Delta\sigma = 1{\cdot}4\,\mathrm{kg/mm^2}$ is again the most suitable one to choose, and gives $vN_0 = 1{\cdot}7 \times 10^{-11}$.

We cannot evaluate either v or N_0 separately without some further assumption and any further discussion must be somewhat speculative. It will be recalled that $N d\sigma$ was defined as the number of dislocation loops per unit volume with activation stresses in the range $d\sigma$. Now the total number of dislocation lines per unit area in a well annealed crystal has been estimated as being of the order of 10^8. Accepting this value for the moment, we obtain

$$\int_0^\infty N\,d\sigma = 10^8/\Lambda\,;$$

using the value of Λ already found, this is of the order $10^{14}/\mathrm{cm}^3$. On our assumption that N is independent of σ, this must be set equal to $N_0(\sigma_{\max} - \sigma_{\min})$ where $\sigma_{\min}$ is the least applied stress that will move a dislocation loop (of the order of the yield stress), and $\sigma_{\max}$ is the stress required to move the most tightly bound dislocation loop. $\sigma_{\max}$ must exist, if only for the reason that the material will not support infinite internal stresses; let us suppose that it lies between 2 and 11 times the yield stress. Then $(\sigma_{\max} - \sigma_{\min}) = \sigma_{\min}$ (or $10\sigma_{\min}$ on the second assumption), i.e. is of the order 10^9 (10^{10}) dyne/cm^2. Thus $N_0 = 10^5$ (10^4) and this would suggest $v = 10^{-16}$ (10^{-15}) cm^3. Now v/V was defined as the strain resulting from the activation of a single dislocation loop, where V is the volume of the crystal. The movement of the activated loop itself is equivalent to slip of one atomic spacing over an area of slip-plane of order of magnitude Λ^2, and thus corresponds to a strain of $a\Lambda^2/V$. Thus if the activated loop is the only one that moves, we have $v = a\Lambda^2$, which is of order 10^{-20} cm^3. Comparing the two figures we may deduce that the activation of a single dislocation loop starts off an avalanche which, on the average, involves about 10^4 (10^5) other similar loops.

Although the preceding calculation is very rough it does suggest that the elementary slip process with which the theory ought to concern itself is one involving a considerable number of dislocation loops. It may be that this will provide a clue to the major problem, that of reconciling the theoretical and experimental estimates of the frequency of vibration of a dislocation loop. Until this is done the theory cannot be considered to be entirely satisfactory. It is also clearly desirable to extend the measurements to other 'hard' materials, to which the theory should apply equally well; it is hoped to do this in due course.

ACKNOWLEDGMENTS

Our best thanks are due to Professor Mott for suggesting this problem and to many colleagues in the H. H. Wills Physical Laboratory for assistance with the work and discussion of the results. One of us (M. D.) also wishes to acknowledge financial assistance from the Ministry of Education under the Armed Forces Further Education Scheme.

REFERENCES

ANDRADE, E. N. DA C., 1910, *Proc. Roy. Soc.* A, **84**, 1; 1914, *Ibid.*, **90**, 329.
MOTT, N. F., and NABARRO, F. R. N., 1948, *Report on Bristol Conference on Strength of Solids* (London: Physical Society), p. 1.
TAYLOR, G. I., 1934, *Proc. Roy. Soc.* A, **145**, 362.

A Study of Electrical Forming Phenomena at Selenium Contacts

By H. K. HENISCH and J. EWELS
Physics Department, University of Reading

Communicated by R. W. Ditchburn; MS. received 13*th January* 1950

ABSTRACT. An account is given of current-creep experiments at various temperatures on Se specimens of different impurity content. The dependence of these phenomena on the electrical and thermal history of the specimens and on the nature of the counter-electrode is investigated. It is shown that two opposing creep mechanisms are, in general, active simultaneously. One is due to power dissipation within the barrier layer, the other due to structural changes which take place under the influence of the applied field. It was found, contrary to theoretical expectations, that creep processes cannot be 'frozen out' at low temperatures (e.g. $-183°$ c.). Measurements of the self-capacitance of the barrier were carried out at suitable stages of the experiment. These show that the thickness of the barrier increases slightly during forming. A theory of current creep is proposed on the basis of the present observations.

§ 1. INTRODUCTION

When a reverse voltage is applied to a rectifying contact between a selenium surface and a metal electrode, the resulting current is not constant but is generally a function of time. An increase of current with time is called positive creep and a decrease at constant voltage is referred to as negative creep. These phenomena, which can also be observed on other semiconducting materials, are the basis of the forming process which is of great practical importance in the manufacture of disc rectifiers, crystal diodes and crystal triodes (transistors). The forming process as applied to selenium rectifiers results in a very considerable improvement of the rectification properties. Junctions which have been 'electrically formed' have a higher D.C. resistance in the blocking direction and are therefore able to withstand a much higher reverse voltage without breakdown than unformed discs. It must be inferred that some structural modification of the barrier layer takes place under the influence of the applied field which may be as high as 10^6 volts/cm. At normal temperatures a formed rectifier retains its improved characteristics, except for the slow deterioration known as ageing. The structural modification of the barrier must therefore be ascribed to a process which is only slightly reversible under these conditions. Its exact mechanism is evidently of great interest, not only because of its technical importance, but because no theory of barrier layer rectification can be complete unless it can account for creep phenomena.

A good deal of information on these matters is probably in the hands of rectifier manufacturers, but very little has appeared in the technical press. Williams and Thompson (1941) give a brief description of forming phenomena on commercial rectifiers. Their results show the complicated nature of the processes involved and their dependence on the thermal and electrical history of the specimens. Rose and Schmidt (1947) have studied similar phenomena on selenium contacts by the application of voltage pulses lasting about one-fifth of a second, and claim to have shown that positive creep is due to an increasing concentration of impurity centres in the immediate neighbourhood of the counter-electrode, and negative creep due to the subsequent migration of centres from the semiconductor into the metal under the influence of the applied field. More recently, Kobayashi (1949) has published a short account of creep phenomena in cuprous oxide rectifiers. In the course of his investigations, the time constant of negative creep was found to be about 2 minutes.

The purpose of the experiments described in the present paper was to obtain a more comprehensive series of results, not on commercial rectifiers whose structure is rather complicated, but on selenium contacts prepared under controlled laboratory conditions *. Amorphous selenium, either pure or with deliberate additives, was deposited on nickel-plated steel discs, as during the manufacture of normal rectifiers, and subsequently heat-treated to convert it to the crystalline form. The counter-electrodes used were either gold (applied by vacuum evaporation) or a low melting-point alloy of 72 % Sn and 28 % Cd (applied by spraying). The experiments consisted mainly in the continuous observation of the currents resulting from the application of various constant reverse voltages to selenium contacts. At convenient intervals the self-capacitance of the contacts at zero bias voltage was determined by measuring the resistive and reactive components of the contact impedance at different frequencies, as described by Henisch (1949). The influence of temperature, electrode material, interruptions in the forming process and impurity content of the selenium on the capacitance and the current–time relation was also studied. The systematic variation of these parameters has led to results by which existing theories of current creep can be tested. It is also possible to obtain a clearer picture of the conditions within the barrier, particularly as regards the relations between thermal effects and ionic diffusion.

§ 2. EXPERIMENTAL TECHNIQUE

The simple circuit used for the present investigation is shown in Figure 1. The reverse voltages applied to the contacts under test were taken from a series of 2 v. accumulators. A potential difference developed across the input coil of the magnetic amplifier and the resulting output current was passed to a recording milliammeter (5 ma. full scale deflection, chart speed 1 in. per min.). The input resistance of the amplifier (6·61 ohms) could be shunted so as to provide various current ranges (e.g. 0–6 ma., 0–30 ma., 0–120 ma., 0–1 amp.). A valve-controlled overload cut-out was installed to protect the current recorder in case of electrical breakdown of the specimen under test. The accumulators were periodically charged *in situ* and before use partly discharged through a dummy load in order to ensure constancy of the applied voltage. Owing to the inertia of the current

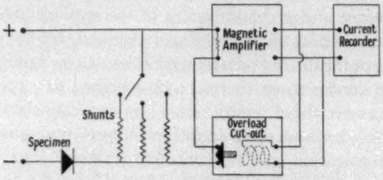


Figure 1. Circuit for current-creep measurements.

---

* Supplied by Messrs. Standard Telecommunication Laboratories Ltd. to the authors specification.

recorder no observations were possible during the first second or so after application of the voltage.

The bridge circuit used for capacitance measurements is shown in Figure 2. The decades in the balancing arm could be arranged cither as a series or a parallel combination, whichever proved more convenient for a particular specimen. The applied alternating voltages never exceeded 20 mv.

The specimens consisted of discs of 3·5 cm. diameter which carried six independent electrodes as shown in Figure 3. One or other of these contacts could be brought into the circuit of Figure 1 by means of a selector switch. During the experiments the discs were mounted in a tubular brass container which could be evacuated and heated (electrically) or cooled (for instance, by means of solid CO_2) from the outside. The ambient temperature was measured by means of a chromel-alumel thermocouple at the centre of the disc. The temperature of the counter-electrode may be somewhat higher, depending on the power dissipated within the barrier layer. During one series of experiments the resulting temperature rise was approximately assessed by placing a thermojunction in direct contact with the counter-electrode. All current leads were taken out of the container through hermetic seals which were electrically heated to prevent frosting over during the low temperature experiments.

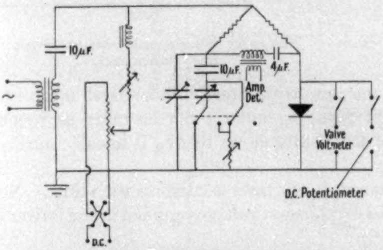


Figure 2. Bridge circuit for the measurement of barrier self-capacitance.

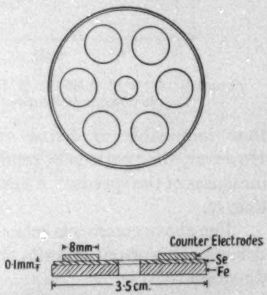


Figure 3. Experimental specimens.

§ 3. GENERAL CHARACTER OF CURRENT-CREEP PHENOMENA

The general creep character of the observed current creep is shown in Figure 4 (curve *a*). The application of a constant reverse voltage to the contact is followed by a short period of positive creep and, in due course, by a much longer period of negative creep. It is reasonable to regard a curve of this kind as the composition of two opposing tendencies, representing pure positive creep (curve *b*), and pure negative creep (curve *c*) respectively. There is no physical reason for expecting these components to be simply exponential. The usefulness of this analysis depends ultimately on the possibility of showing that two distinct—though not necessarily wholly independent—physical processes are at work. Evidence for this conclusion will be given and discussed in the following sections.

In Figure 4, the speeds of the two processes and their amplitudes are assumed to be comparable. In general this need not be so. Hence, while opposing tendencies may both be operative, the current maximum may or may not be observed, depending on the relative magnitude of j_H and j_F as well as the speeds of the corresponding processes. The current–time relation may thus appear in a number of forms, examples of which are shown in Figure 5. Positive or negative creep may be suppressed or, in a special case, the two processes may balance initialiy (e.g. as in curve *c*) so as to keep the current temporarily constant. Simple tests show that curves of type *a* and *b* are indeed not simply exponential and it is thus impossible to define time constants with precise theoretical meaning. However, the 'periods of semi-completion' (t_H and t_F) can be taken as rough measures of the speeds. A review of the results shows that t_H is usually smaller than t_F.

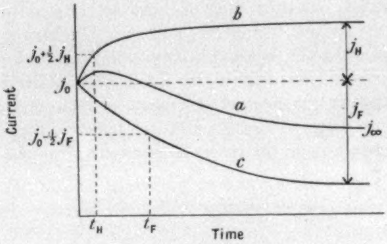


Figure 4. General character of current creep phenomena (schematic).

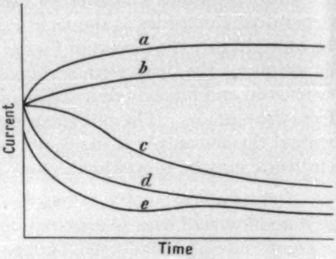


Figure 5. Some special cases of current creep (schematic).

The above comments relate to changes in the reverse currents with time. No current creep of any kind was observed for constant voltages applied in the forward direction.

§ 4. INFLUENCE OF ELECTRICAL HISTORY

For the sake of simplicity and in order to establish a basis for comparison, the discussion in this and the two following sections will be restricted to experimental results obtained on discs with alloy counter-electrodes on selenium containing 0·02 % of iodine. In order to assess the influence of electrical history, five types of current-creep experiments were carried out:

(*a*) formation of a rectifying junction at successively increasing reverse voltages up to a safe maximum, insufficient to cause breakdown;

(*b*) formation of a rectifying junction at a single high reverse voltage;

(*c*) application of a low reverse voltage shortly after formation of the barrier at a higher voltage;

(*d*) repeated application of the highest previous forming voltage at short intervals;

(*e*) repeated interruption of the forming process before the establishment of equilibrium, i.e. before a constant current has been reached.

Figure 6 shows the progress of current creep on two new junctions at reverse voltages increasing in steps of about 4 volts, with 10-minute intervals between successive stages. Capacitance measurements were carried out before this sequence and during the intervals. In order to save space, only the first six minutes of the creep record are reproduced. Where applicable, the current at the conclusion of the run and the corresponding time are indicated on the right-hand side of the graphs. It should be noted that the scale and origin on the current axis is not the same for all curves. Negative creep is observed in every case, but the initial positive creep (and hence the current maximum) only for higher voltages. Different electrodes on the same disc behave in an essentially similar way, though the values of j_0 and j_∞ may vary somewhat from one junction to another.

These results are typical for current creep at room temperature. Experiments carried out at different temperatures—and, of course, on different junctions—are not strictly comparable in view of the normal variations between different electrodes, and because the effect of temperature changes is not perfectly reversible. Even in absence of an externally applied voltage, the barrier carries the normal diffusion field under the influence of which almost irreversible modifications in the structure of the barrier may take place. Experiments carried out at low temperatures (e.g. $-59°$ c.) and high temperatures (e.g. $+60°$ c.) do not reveal any characteristic deviations from the relations shown in Figure 6. Creep phenomena are still very pronounced at the lower temperatures and are not markedly enhanced at $+60°$ c. as compared with room temperature. Even at liquid air temperature negative creep is still clearly observable though no positive creep was encountered during the short experimental series in this range.

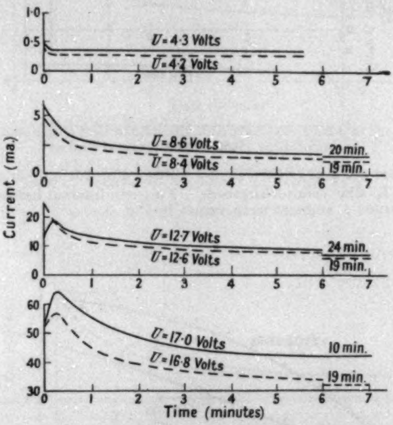


Figure 6. Record of creep phenomena during forming at successively increasing reverse voltages (U).

Se+0·02% I_2 alloy counter-electrodes.
10 minute intervals between runs.
—— Disc *b*, electrode 1, ambient temperature 22° c.
– – – Disc *b*, electrode 2, ambient temperature 19° c.

If, as an alternative to forming in stages, a high reverse voltage is directly applied to a new junction, the final resistances are not very different from those observed after forming in stages. They are reached after pronounced negative creep. The application of a comparatively low reverse voltage shortly after forming at a higher voltage results in positive creep only. An example of this is shown in Figure 7. The same is true if the highest previous forming voltage is repeatedly re-applied after short intervals, i.e. intervals not exceeding 1 hour or so. (Phenomena observed after longer intervals will be dealt with in the next section.) Positive creep occurs in every case, to an extent which depends on the length of the preceding interval. Figure 8 shows how $j_\infty - j_0 (= j_H)$ varies with the length of this interval, for three different junctions. Such regular characteristics of successive creep curves on a given junction have never been observed in connection with negative creep. They are associated with positive creep only, a fact which supports the hypothesis that positive and negative creep are due to different mechanisms. This is confirmed by the form of the creep curve obtained when the forming process is repeatedly interrupted before equilibrium is reached, as shown in Figure 9. It will be noted that every short interruption (1 second or so) is followed by some positive creep, after which the original negative creep is resumed in much the same form as that in which it is known normally to proceed when the applied voltage is constant (broken line). If, on the other hand, the duration of the interruption becomes comparable with the total creep period, the current–time relation is as shown in Figure 10. The shape of the curve is the same as before, but a

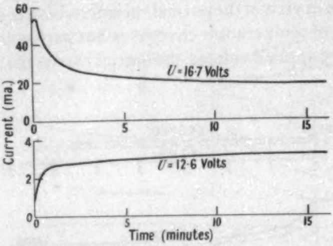


Figure 7. Current creep at a relatively low voltage after forming at a higher voltage. Se+0·02% I_2 alloy counter-electrode. 5 minute interval between runs. Disc *e*, electrode 4, ambient temperature 19·5° c.

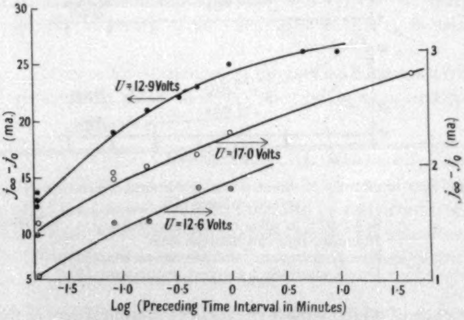


Figure 8. Current creep after re-application of the highest previous forming voltage. Relation between $j_\infty - j_0 (= j_H)$ and preceding time interval.
Se+0·02% I_2 alloy counter-electrode.
○ Disc *b*, electrode 1, ambient temperature 22° c.
● Disc *b*, electrode 3, ambient temperature 17° c.
⊙ Disc *c*, electrode 2, ambient temperature −59° c.

displacement of the second part towards the right by an amount which corresponds to the period of interruption can now be observed. The broken line shows approximately the manner in which creep was expected to proceed if the applied voltage had remained constant. It appears that after the re-application of the external voltage, negative creep is resumed, but that its initial phases are more than compensated by positive creep.

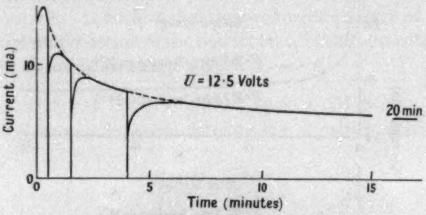


Figure 9. Current creep following short interruptions (1 sec.) of the forming process before the establishment of equilibrium.
Se+0·02% I_2 alloy counter-electrode.
Disc *e*, electrode 6, ambient temperature 21° c.

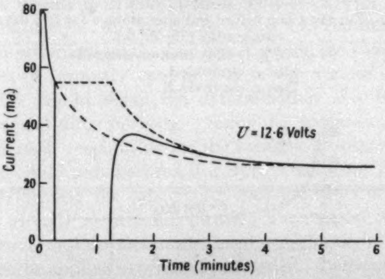


Figure 10. Current creep following a long interruption (approximately 1 minute) of the forming process before the establishment of equilibrium.
Se+0·02% I_2 alloy counter-electrode.
Disc *g*, electrode 2, ambient temperature 20° c.

§ 5. STABILITY

When a junction has undergone a series of electrical operations of the type shown in Figure 6, the barrier has not reached a perfectly stable condition. It is subject to further changes during storage, even in absence of an applied forming voltage. These changes are dependent on the storage temperature and can also be influenced by the passage of large direct currents in the forward direction. Both factors were investigated by observing original and final current-creep curves for a given applied voltage.

If a junction which has previously been formed at a high reverse voltage is stored for two days at room temperature and subsequently re-formed at the same voltage, the resulting creep curve is very similar to that observed before the storage period, but negative creep is less prominent. The final currents differ only slightly from the values of j_∞ before storage. On the other hand, if similar measurements are carried out at a high temperature (e.g. 75 to 80° c.) the final currents are considerably lower and are reached after an additional negative creep process. Figure 11 illustrates this point. Storage at high temperatures thus displaces the equilibrium previously reached at room temperature and allows a further modification of the barrier in the same direction.

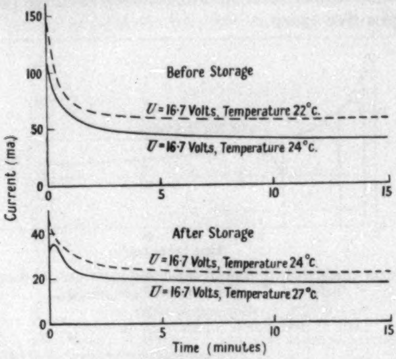
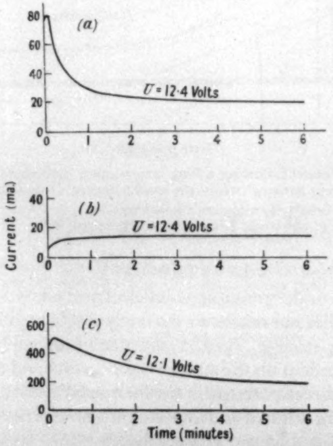


Figure 11. Current creep before and after storage for two days at a high temperature (75–80° c.).
Se+0·02% I_2 alloy counter-electrode.
—— Disc *e*, electrode 1.
--- Disc *e*, electrode 2.



(*a*) First application of forming voltage before passage of forward current.
(*b*) Re-application of forming voltage before passage of forward current.
(*c*) Re-application of forming voltage after passage of forward current.

Figure 12. Current creep before and after the passage of forward current (200 ma. for 45 hr.).
Se+0·02% I_2 alloy counter-electrode.
Disc *g*, electrode 6, ambient temperature 18° c.

The influence of forward current on the properties of the rectifying junction is shown in Figure 12. The first curve represents the original forming process at 12·4 volts. After an interval of 1 hour the same voltage was re-applied and the second curve shows the corresponding creep relation which is of the type already discussed in the previous chapter. A forward current of 200 ma. was then passed for 45 hours, and re-application of the original reverse voltage then resulted in current creep as shown by the third curve. It should be noted that j_∞ has increased by a factor of about 10. It is clear that the prolonged passage of forward current causes a substantial modification of the barrier layer in the opposite direction to the change due to the forming process.

§ 6. VARIATION OF SELF-CAPACITANCE DURING FORMING

Schottky's well-known theoretical treatment of self-capacitance leads to the expression

$$\frac{d(1/c^2)}{dU}=\frac{8\pi}{K\mathbf{e}N_\lambda},$$

where U is the applied voltage (direct), K the dielectric constant, $\mathbf{e}$ the electronic charge, N_λ the concentration of impurity centres at a distance λ from the metal boundary, and c the dynamic capacitance (dQ/dU) per unit area of the barrier layer. The above equation in conjunction with $c=K/4\pi\lambda$, can be used for a graphical evaluation of N_λ for any particular value of λ (Henisch 1949). This was thought to offer an attractive method of investigating the changes (if any) in the local concentration of impurity centres due to the various stages of forming. Such measurements would entail the determination of c for various applied voltages U, before and after forming. Since the application of these voltages to new junctions is itself responsible for the forming process, it was realized that Schottky's method could only be used if it were possible to 'freeze out' all creep phenomena by performing the capacitance measurements at a sufficiently low temperature. As already mentioned, forming phenomena are still observed at liquid air temperatures and the technique based on Schottky's capacitance relation could not, therefore, be applied as originally intended. In the cases here described, its usefulness was limited to the small range of reverse voltages (e.g. $U<3$ volts) for which current creep is negligible. (In the forward direction the measurement of self-capacitance is more difficult owing to the low shunt resistance.) Figure 13 shows $1/c^2$ plotted against U. The resulting relation is a straight line which on the basis of Schottky's treatment implies a constant concentration of impurity centres. The value of N deduced from the slope is approximately 7×10^{16}. Some variations from one disc to another were observed, but this concentration can be taken as typical. The intercept on the abscissa should be numerically equal to the diffusion potential and is found to be about 0·6 volt. Unfortunately, the slope of such lines is not a sensitive measure of N, but the results indicate that forming produces a small increase in the concentration of impurity centres at distances between 2×10^{-5} cm. and 4×10^{-5} cm. from the counter-electrode.

Capacitance measurements in absence of a direct bias voltage are not, of course, subject to the above disturbances and represent a useful, though less informative, method of detecting changes in the structure of barrier layers. Such measurements were carried out after each stage of forming, and typical results are shown in Figure 14. It will be noted that the forming process is accompanied by a small decrease of self-capacitance, corresponding (as expected) to a thickening of the barrier layer. The capacitances decrease with decreasing temperature. The series of measurements shown in Figure 14 were, of course, carried out on different junctions, but the self-capacitances determined at low temperatures were quite generally found to be smaller than those measured at high temperatures. The width of the barrier at normal temperatures is about 2×10^{-5} cm.

§ 7. EFFECT OF COUNTER-ELECTRODE MATERIAL AND OF ADDITIVES TO THE SELENIUM

Experiments of the kind described in the previous sections were also carried out on specimens with counter-electrodes of (vacuum-evaporated) gold, on selenium with and without deliberate additives. The results are summarized by Table 1 which gives the initial and final currents for two forming voltages.

Comparing values for j_0, it will be seen that counter-electrodes of gold are always associated with lower barrier resistances than the corresponding alloy electrodes. This applies at both forming voltages. Table 2 gives the mean values of $(j_0 - j_\infty)/j_0$ (averaged over the whole temperature range) and the corresponding mean deviations for the various groups. These values show that the forming process is always less effective for gold than for the Sn–Cd alloy. The results also indicate that the effect of the additives to the selenium is remarkably small. This applies also to the current–time relation during forming, to the dependence of the forming process on the electrical history of the junction and to the self-capacitance. Iodine causes a small decrease of barrier resistance, and mercury (added before melting) a slight increase. Contrary to expectation, the initial and final currents of low-temperature runs are nearly always higher than those for room temperature and above.

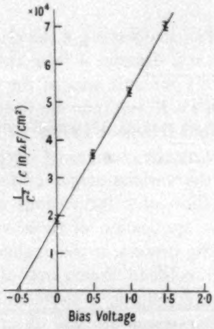


Figure 13. Schottky's capacitance relation plotted for disc *e*, electrode 5 after forming at 4 volts.

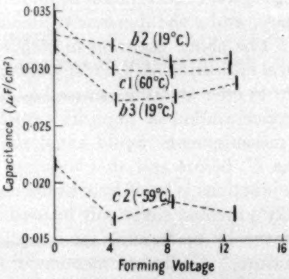


Figure 14. Changes of barrier capacitance due to various stages of forming. Se+0·02% I_2 alloy counter-electrode.

The effect of mercury is quite different when specimens are exposed to mercury vapour. After short periods (e.g. several days in the presence of air at atmospheric pressure) the barrier resistance falls to a small fraction of its original value and the junction becomes useless for rectification purposes. This phenomenon represents a well-known danger to industrial rectification plant. The results

Table 1

Summary of Current Creep Experiments on Different Types of Junction

Counter-electrode	Additive	Junction No.*	Temp. (°c.)	Forming at 8·2 volts		Forming at 12·5 volts	
				j_0 (ma.)	j_∞ (ma.)	j_0 (ma.)	j_∞ (ma.)
Alloy	Nil	b1	+20	1·20	0·35	4·9	1·05
		b2	+22	0·65	0·30	4·0	1·20
		c2	+68	0·87	0·25	3·1	0·56
		c4	−54	6·2	1·2	27·0	4·2
		c5	−57	7·5	1·2	29·7	2·7
Gold	Nil	a2	+20	190	128	193	182
		a1	+23	109	99	—	—
		b2	+70	59	32	97	43
		b1	+70	80	42	—	—
		b5	−48	196	175	—	—
		b4	−50	262	210	—	—
Alloy	0·02% I_2	b1	+22	4·9	1·6	16·0	7·0
		b2	+19	4·0	1·2	24·0	6·2
		b3	+17	4·2	2·6	22·0	7·5
		c1	+60	4·4	1·2	14·4	4·2
		c6	+60	1·6	0·8	5·0	3·1
		c3	−58	9·5	1·7	31·0	6·5
		c2	−59	9·2	1·7	42·0	6·6
Gold	0·02% I_2	a3	+25	125	102	160	135
		a2	+30	140	109	—	—
		b2	+61	141	89	82	83
		b1	+61	170	95	—	—
		c2	−56	245	180	275	185
		c1	−57	38	(42†)	71	62
Alloy	0·05 %Hg‡	a3	+19	1·47	0·72	5·5	2·85
		a6	+65	0·78	0·39	3·3	0·97
		a5	+65	0·95	0·45	2·7	1·00
		a1	−49	9·6	2·30	35	5·5
		a2	−50	7·0	2·0	37	4·2
Gold	0·05% Hg‡	a2	+27	77	48	118	74
		a1	+22	83	53	129	71
		a6	+65	35	12·5	41	23·5
		a3	+60	55	30	88	44
		a4	−56	98	65	—	—
Alloy	0·02% I_2 0·05% Hg‡	a2	+21	4·3	0·82	16·0	4·2

* The letters refer to different discs, the numbers to individual counter-electrodes (see Figure 3).

† Positive creep only.

‡ Added before melting.

previously quoted show that mercury is not harmful when added to the selenium before melting. After melting and crystallization it is then, presumably, present in interstitial positions or in some chemical combination, whereas exposure of the crystalline material to mercury vapour is more likely to result in adsorption at grain boundaries.

Table 2. Effectiveness of the Forming Process (at 8·2 volts) for Different Types of Junction

Counter-electrode	Additive	Average value of $(j_0 - j_\infty)/j_0$	Mean deviation
Alloy	Nil	0·72	0·08
Gold	Nil	0·31	0·14
Alloy	0·02% I_2	0·66	0·13
Gold	0·02% I_2	0·30	0·09
Alloy	0·05% Hg	0·60	0·11
Gold	0·05% Hg	0·43	0·09

§ 8. THE MECHANISM OF POSITIVE CREEP

The results described in the previous sections show that there are characteristic differences between the conditions for the occurrence of positive and negative creep. Thus negative creep is a process which depends critically on the previous history of the junctions and is generally not repeatable. Positive creep, on the other hand, is well reproducible when sufficiently high voltages are applied to junctions which are fully formed. It was therefore concluded that positive and negative creep are due to entirely different mechanisms, and that the observations of current creep under various conditions cannot be simply accounted for on the basis of existing theories.

The manner in which these differences manifest themselves indicates that positive creep is not caused by structural changes, but is due to the power dissipation within the barrier layer. This conclusion is supported by the fact that positive creep is always prominent when the power dissipation is high, and by the repeatability of the phenomenon. In order to examine this interpretation in more detail, a special series of measurements was carried out during which the temperature at the top of the counter-electrode (Figure 3) was repeatedly measured during negative and positive creep. Typical results are shown in Figure 15.

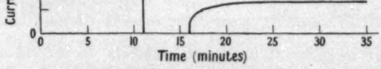


Figure 15. Record of counter-electrode temperature during an interrupted creep process.
Se+0·02% I_2 alloy counter-electrode.
Disc *a*, electrode 5.

The first application of a high reverse voltage results, as usual, in pronounced negative creep and also in some heating of the barrier layer. The actual temperature of the barrier layer itself is probably considerably higher than that given on the graph. This temperature rise counteracts the forming process and after about two minutes (in this case) the two opposing tendencies balance. However, since temperature stability is not yet reached at this stage, the current thereafter increases, though more slowly. When the temperature of the barrier has become constant, slight negative creep may take over again until the forming process is completed. If the reverse voltage is then switched off, the counter-electrode cools down, as expected. Subsequent re-application of the same voltage results in prominent positive creep, followed by only slight negative creep, since the forming process was almost completed during the previous application. The corresponding temperature rise is again shown in Figure 15, and can always be observed whenever positive creep occurs. The second part of the graph in Figure 15 can be repeated many times without any appreciable change of shape.

On the above scheme one would expect that the re-application of a forming voltage after a long interruption should give rise to pronounced positive creep, whereas a short interruption during which there is little opportunity for cooling should be followed by only slight positive creep. This is confirmed by the results plotted in Figure 8.

The above interpretation implies that a rise in the temperature of the barrier necessarily results in a lowering of the barrier resistance. This has been confirmed by potential probe measurements on single barriers. The results in Table 1 show, however, that for the rectifier assembly as a whole, the resistance per unit area decreases as the temperature decreases. This behaviour has also been observed on previous occasions (U.S. Signal Corps Engineering Laboratories 1949) and its interpretation represents a major problem, independent of any considerations of rectifier forming. In some way the effect of heating the assembly as a whole is not the same as that of heating only the barrier layer and its immediate surroundings. Capacitance measurements have not revealed the existence of any resistance in series with the barrier which could be made responsible for the unexpected temperature dependence, even if a suitable mechanism could be postulated. It seems possible, therefore, that this behaviour is due to a change in the effective contact area with temperature. The manner in which such a change could be brought about is still uncertain. It may, conceivably, be due to differential expansion at the base electrode or in some other way due to lateral variations in the structure of the system. This matter will be referred to again in a future communication.

§ 9. THE MECHANISM OF NEGATIVE CREEP

An interpretation of negative creep must take into account the following observations:

(*a*) Negative creep depends critically on the previous electrical and thermal history of the rectifying junction.

(*b*) The effect of forming is only slightly reversible under normal storage conditions, but the prolonged passage of forward current results in a considerable amount of de-forming.

(*c*) Storage of a fully formed junction at a high temperature (e.g. 75–80° c.) establishes conditions under which a further forming process is possible.

(*d*) When allowance is made for an initial positive creep due to power dissipation, it is found that a forming process can be interrupted repeatedly without affecting the manner of its subsequent completion (Figures 9 and 10).

(*e*) Negative creep can still take place at low temperatures (e.g. −183° c.).

(*f*) Negative creep is not noticeably affected by the presence of considerable amounts of iodine and mercury (added before melting) in the selenium.

(*g*) The forming process is more effective for counter-electrodes of Sn–Cd alloy than for those of gold. Counter-electrodes of gold also have the lower barrier resistance.

(*h*) The self-capacitance of the barrier layer decreases during the forming process.

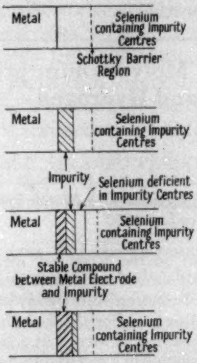
(*i*) A test on the basis of Schottky's capacitance method (as far as it is applicable) does not reveal large gradients in the concentration of impurity centres near the counter-electrode before or after forming.

Selenium is known to be a deficit semiconductor from the polarity of its thermoelectric power and from the direction of rectification. Its impurity centres thus have an electro-negative character. They are presumably attracted towards the counter-electrode by the normal diffusion field, and more strongly attracted when a reverse voltage is applied to the barrier layer. Since the field strengths involved are large, it is to be expected that these impurity centres will, in the course of time, be displaced towards the counter-electrode until a new equilibrium is established. In absence of an external field the same process should occur, though more slowly and not to the same stage of completion, i.e. the junction is expected to undergo a partial self-forming process before any reverse voltage is applied. Junctions which have counter-electrodes of apparently equal area are frequently found to have different resistances (per unit area) although prepared by identical methods. A self-forming process of the type here envisaged would account for such differences.

The diffusion field is always strongest in the immediate proximity of the counter-electrode and this is the region from which impurity centres are most likely to be displaced. The exact nature of the centres is still unknown. Since iodine, which is often added to selenium in the course of rectifier production, does not materially affect the creep process, it must be concluded that it does not form active impurity centres. Mercury (which is in any case more likely to be an electron donor than an acceptor) likewise does not appear to form such centres. These conclusions are supported by thermoelectric measurements on selenium which will be described in another paper.

What function can be ascribed to the displaced impurity centres (whatever their exact nature may eventually be found to be)? It is possible to envisage that these centres may separate out at the interface between the selenium and the counter-electrode, or may combine with the counter-electrode to form a stable compound provided the electrode is made of a suitable material. A chemical barrier layer would result in both cases. The normally good stability of fully formed rectifiers lends probability to the second hypothesis. A counter-electrode of Sn–Cd alloy would then behave rather differently from one of gold which is chemically almost inert. Such differences of behaviour have already been described. The gold counter-electrodes are associated with lower barrier resistances and a less effective forming process, though this may be partly due to differences in the effective contact areas. The two mechanisms for forming here suggested are not mutually exclusive. The separation of impurity centres may, and probably does, precede the formation of a compound with the counter-electrode. Figure 16 shows schematically the various stages of the forming process as envisaged by the present theory.

Once an impurity ion has entered into combination with the counter-electrode, it is firmly held, but impurities which have merely separated out and have not yet entered into combination may, under suitable conditions, diffuse back into the selenium. Conditions would be particularly favourable for this when a forward voltage is applied. This would account for the observed de-forming produced by the prolonged passage of forward current through the junction. In the absence of a forward voltage the same process would take place, though again more slowly. This is probably the mechanism of what is commercially called 'ageing'.


Metal
Selenium containing Impurity Centres
Schottky Barrier Region
Metal
Selenium containing Impurity Centres
Impurity
Selenium deficient in Impurity Centres
Metal
Selenium containing Impurity Centres
Stable Compound between Metal Electrode and Impurity
Metal
Selenium containing Impurity Centres

(*a*) Junction immediately after application of counter electrode.

(*b*) Junction after prolonged self-forming under the influence of the diffusion field.

(*c*) Junction immediately after the forming process.

(*d*) Formed junction after ageing (scope for a new forming process).

Figure 16. Schematic representation of the forming process as suggested by current creep measurements.

It is now necessary to consider how the forming process ends. On the basis of the present model, negative creep would be expected to cease eventually for two reasons: (*a*) because the field strength within the selenium decreases as the chemical barrier develops, and (*b*) because the region in the selenium nearest to the counter-electrode becomes exhausted of impurity centres. The experimental results suggest that the second limitation is the more important. During prolonged storage, particularly at higher temperatures, the supply of impurity centres could be replenished by diffusion from the selenium outside the barrier region and perhaps also by re-diffusion of the impurity deposit at the interface (see Figure 16(*c*)). If a high reverse voltage is then re-applied, a further forming process could take place, as shown in Figure 11.

One of the unexpected results is the observation that current creep can still take place at very low temperatures. Ionic motion in a crystal lattice is normally characterized by high activation energies and is therefore expected to be very temperature dependent. The present observations suggest an electrolytic process of remarkably low activation energy, as could only be envisaged if the selenium were in a low state of order, i.e. in a form intermediate between amorphous and crystalline. X-ray measurements tend to corroborate this interpretation.

ACKNOWLEDGMENTS

The authors would like to make grateful acknowledgment to Prof. R. W. Ditchburn for placing research facilities at their disposal and for much valuable help and advice. Thanks are also due to Mr. T. R. Scott of Standard Telecommunication Laboratories Ltd. for his interest in this work and for kindly supplying the test specimens.

REFERENCES

Henisch, H. K., 1949, *Metal Rectifiers* (Oxford : Clarendon Press).
Kobayashi, A., 1949, *J. Phys. Soc., Japan*, **3,** 41.
Rose, F., and Schmidt, H., 1947, *Z. Naturforschung*, **2a,** 226.
Williams, A. L., and Thompson, L. E., 1941, *J. Instn. Elect. Engrs.*, Pt. I, **88,** 353.
U.S.A. Signal Corps Engineering Laboratories, 1949, *Technical Memorandum* M-1173, Fort Monmouth, N.J., U.S.A., 31st January 1949.

Secondary Electron Emission from Molybdenum produced by Helium, Neon, Argon and Hydrogen

By D. GREENE
Department of Experimental Physics, Queen's University, Belfast

Communicated by K. G. Emeléus; MS. received 13*th March* 1950
To be read on 24*th November*, 1950

ABSTRACT. Experiments have been carried out on the secondary electron emission produced at very low pressure from a degassed metal surface by neutral particles resulting from the neutralization of positive ions of about 1,000 electron volts energy in a metal canal. The gases used were helium, neon, argon and hydrogen. For the first three the energy distribution of the secondary electrons is compatible with the view that a large part of the emission is due to communication of the energy of excitation of a metastable atom to a metallic electron. With hydrogen there is evidence that a similar process occurs with normally excited states of the atom or molecule. In all cases there is evidence that some emission occurs through local heating or disruption of the metal lattice by the mechanical impact of swift atoms, the extent of this increasing from helium and neon to argon and hydrogen. Some reflection of the swift atoms occurs from the metal target.

INTRODUCTION

The secondary electron emission produced by metastable helium atoms at a degassed metal surface in a vacuum has been investigated by Oliphant (1929). He found that the energy distribution of the electrons had an upper limit, equal to the energy of excitation of the metastable atom, less the work function of the metal, and a lower limit, which agreed well with the value of the energy of excitation of the metastable atom less a calculated value for the total potential barrier at the surface of the metal. These observations seemed to show that the secondary electron emission could be regarded as the result of individual interactions between the incident metastable atoms, and the electrons in the metal. The theory of the energy distribution amongst the secondary electrons ejected in this way has been developed by Massey (1930, 1931). In Massey's theory the metal is treated as a molecule whose ionization potential is equal to the work function, and the interaction considered as a collision of the second kind. The theoretical energy distribution shows a higher proportion of high energy electrons than was found by Oliphant, and a sharper cut-off at the upper energy limit. Massey expressed the opinion, however, that the agreement with experiment was as good as could be expected from the approximate nature of the calculations and the difficulties of the experimental work.

Chaudhri and Khan (1948) have studied the secondary emission produced by neutral atoms or molecules of mercury and potassium. Although mercury has metastable states, their excitation energy is little greater than the work function of the target, so that the source of energy for individual interactions of the kind considered by Massey is small. Potassium has no metastable state and therefore no internal energy available for such interactions, under the conditions of experiment. Chaudhri and Khan found that the upper electron energy limit for both potassium and mercury increased as the kinetic energy of the incident particles was increased, and that a semi-logarithmic plot of current of secondary electrons to a collector against retarding voltage gave nearly a straight line over a range of several volts. They interpreted this as showing that the emission was caused by highly localized heating of the target, due to transfer of the kinetic energy of the particles of the incident beam. From the slope of the line obtained by plotting the logarithms of the collector current against the collector potential a temperature was deduced for the secondary electrons, which could be taken to be the same as the temperature of the heated spot from which the thermionic emission was occurring. A similar explanation had been given previously of the secondary electron emission produced by fast alkali ions (Oliphant and Moon 1930).

It thus seems that the secondary electron emission may be due to not less than two processes: (i) individual interactions which are similar to collisions of the second kind; (ii) interactions where the kinetic energy of the incident particles causes a localized excitation or disruption of the metallic lattice, resulting in what can be described formally as a thermionic emission. The disruption of the lattice could be due either to the impact of a single particle or to successive impacts. In the former case the electron temperature would be independent of the intensity of the primary beam, while in the latter one would expect the electron temperature to increase with intensity. There is no experimental evidence available on this point but it is thought that the single impact process is the more likely. In this connection it should be noted that Oliphant (1929) observed that sputtering was produced by the beam of metastable helium atoms, showing that some disruption of the lattice did occur; even in this later experiment therefore some of the electron emission observed may thus have been thermionic.

The main purpose of the present experiments was to confirm Oliphant's results for helium, and to extend them to neon and argon. The results which have been obtained are in general agreement with Oliphant's, but more evidence has been found that emission due to localized heating takes place. Some experiments have also been carried out with hydrogen, which have given results in agreement with those of Chaudhri and Khan, and in addition show that electron emission may be produced by impact of atoms or molecules in ordinary excited states.

The experiments of Oliphant, Chaudhri and Khan, and the present author are not directly comparable with others, of which those of Uyterhoeven and Harrington (1930) are typical, where the target surface was located in a discharge plasma, since secondary electron emission is very sensitive to gas contamination of the surface.

§1. APPARATUS AND EXPERIMENTAL PROCEDURE

The final form of the experimental tube is shown in Figure 1. Gas was admitted to the bulb O, where an arc discharge of a few tenths of an ampere was passed between the spiral filament F_1 and the anode A. The electrode P was made strongly negative to the anode so that positive ions were drawn to it, some passing down a metal canal 5 cm. long and 2 mm. in diameter, into the region H, which was maintained at a hard vacuum by a high-speed oil diffusion pump of the type described by Sloan, Thornton and Jenkins (1935). Oil molecules were prevented from diffusing back into H by water-cooled baffles in the throat of the pump and a recessed liquid oxygen trap in the pumping lead. The field, applied between A and P, was reversed between the two coaxial cylinders P_1 and E_1 so that positive ions emerging from the canal were prevented from passing down the tube. Secondary electrons or negative ions produced in the canal would be accelerated to E_1 and, to remove these, the same potential difference as was applied between P_1 and E_1 was applied between the electrodes E_2 and E_3, two smaller semi-cylinders. A, E_1 and E_3 were at the same potential. The target T was of molybdenum, and formed the lid of a pill box which contained the filament F_2. This was insulated from the pill box and could be used to raise the target temperature to bright red heat by radiation and electron bombardment. The cylinder C was used to collect charged particles from the target while E_4, a cylinder containing a central diaphragm, served as a shield for C. E_4 and C were also at the same potential as E_1 and E_3. The electrodes were all made of nickel except the canal P which was aluminium, the tungsten filaments and tungsten leads through the Pyrex glass.

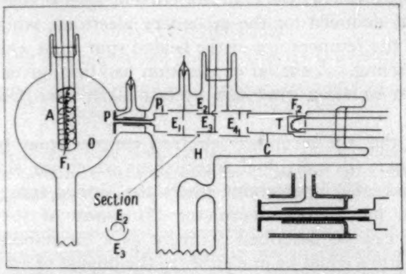


Figure 1. Experimental tube. Inset : details of canal on larger scale.

The potential applied to accelerate positive ions from the arc plasma into the canal was provided by a battery of accumulators of up to 1,500 volts connected between the anode A and the canal P through an 80,000-ohm ballast resistance in the lead to the canal. The applied voltage was measured with an electrostatic voltmeter. All other potential differences were also provided by lead accumulators.

The secondary electron current between target and collector was measured with a galvanometer of sensitivity 5,000 mm/μa. at 1 metre. The total current entering the target–collector system was measured by a galvanometer approximately one-tenth as sensitive. The latter showed zero current during all the secondary electron experiments.

The helium, neon and argon used were obtained from the British Oxygen Company as spectroscopically pure. The hydrogen used was admitted through a palladium leak.

The gas pumped from the high vacuum side of the tube could either be circulated back to O, or pumped out into the air by a mercury diffusion pump backed by a 'Metrovac' mechanical pump. Traps cooled with liquid oxygen were inserted in all the leads to the experimental tube.

Before admitting gas to the system it was evacuated to a pressure of 10^{-6} mm. Hg or less by the auxiliary mercury diffusion pump, and the glass work was torched. The pressure was measured with a McLeod gauge which could be connected to either side of the experimental tube. The target was degassed by maintaining it at bright red heat in the best attainable vacuum. It was found that under these circumstances the thermionic current from the target consisted mainly of positive ions. This emission decreased by a factor of several hundred over a period of about seven hours, after which it reached a steady value of the order of 10^{-8} amp/cm^2. Subsequent reheating of the target, even after it had been exposed to the atmosphere, brought it back to the steady state in about half-an-hour. Before each set of observations on the secondary electron emission the target was degassed in this way for at least this length of time. The target temperature was then reduced to the maximum at which thermionic emission was negligible; this corresponded to a state of dull red heat.

The neon, argon and hydrogen were allowed to stream through the apparatus. The helium was circulated and, to prevent the accumulation of impurities, was continuously purified by passing it through activated charcoal cooled in liquid oxygen before it entered the arc. To avoid contamination of the target, the gas pressure in the target side of the tube must be kept as low as possible. The minimum usable pressure was determined by the pressure required in the arc to give a sufficiently high ion concentration. This was found to vary with the different gases. The upper usable pressure on the high vacuum side of P is the pressure required to give a mean free path equal to the path distance of the neutral particles. These two factors restrict the range of working pressures in O. For neon and argon the main observations were made with a pressure of 8×10^{-5} mm. Hg in the target side of the tube, while for helium and hydrogen the working pressures were 6×10^{-4} mm. Hg and 2×10^{-5} mm. Hg respectively. The corresponding gas pressures in the arc were $10^{-1} - 10^{-2}$ mm. Hg.

In addition to the tube described several others were tried. The details of the electrode structures were different, but the main results obtained were the same in all cases. One of the earlier tubes had a longer neck (5 cm.) than that shown in Figure 1 between the bulb O and the region H. Some results obtained with the long-necked tube will be referred to in § 2.

§ 2. PRELIMINARY OBSERVATIONS

In a preliminary experiment, the electrodes E_1, E_2, E_3 and E_4 were joined to P so that positive ions could reach T and C, which were joined together. The energy distribution of the positive ions was investigated by the application of a variable retarding potential difference between T and E_4. The positive ion current was found to cut off within 3 volts when the retarding voltage approached the accelerating voltage, and a small reverse current which saturated rapidly was observed when the retarding voltage was further increased. This showed (i) that positive ions could not reach the target under the conditions of the main experiment; (ii) that the positive ions coming from the canal had an energy spread of between two and three volts. The small reverse current was presumably due to secondary electrons produced by positive ions colliding with the edges of the slits in the electrodes E_1 and E_4. With the field between E_2 and E_3 as in the main experiments, these secondary electrons could not reach the target. The small energy spread of the positive ions indicates that few gas collisions are taking place in and beyond the canal, and that the particles producing the secondary emission in the main experiments came direct from the arc or resulted from collision between positive ions and the walls of the canal. From the geometry of the tube it is likely that these particles are neutral atoms formed by positive ions capturing electrons from the metal of the canal wall, in making glancing collisions with it. Further evidence for this was found using the long-necked experimental tube (§ 1). On applying a localized magnetic field perpendicular to the axis of the canal, the total secondary electron current from the target was found to vary with the magnetic field strength, when only neutral particles could reach T, in the manner shown in Figure 2. The magnetic field had no effect on the measured arc current or on the appearance of the luminous plasma in O. The explanation of this curve seems to be as follows. As the magnetic field is increased from zero the positive ions passing down the canal have their paths increasingly bent so that the number of collisions with the walls is increased, producing an increase in the number of swift neutral particles emerging into the high vacuum. Further increase in the field bends the paths of the positive ions so much that their collisions with the walls are no longer glancing. This results in a decrease in the number of neutral particles able to emerge and therefore in the secondary electron current from the target. The fact that the decrease to zero of the secondary current is not sharp may be due to the non-uniformity of the magnetic field, which was necessary in order to localize it. The curve was reproducible in its main features. It was not possible to obtain it for the final tube, in which the short neck made it impossible to apply the magnetic field without disturbing the arc.

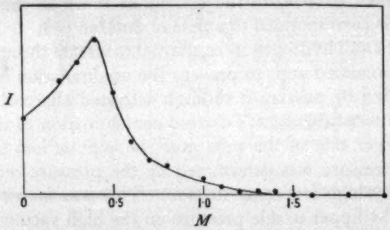


Figure 2. Effect of magnetic field perpendicular to canal; M = magnet current; I = total secondary electron current, arbitrary units. Magnetic field for peak current, approx. 90 gauss.

These results support Oliphant's conclusions that the particles producing the secondary electron emission are metastable atoms formed by glancing collisions of positive ions with the canal walls, and that the secondary emission is due to neutral atoms rather than to quanta coming from the arc. It will, therefore, be assumed that in the present experiments emission due to quanta was negligible for all the gases used, as the experimental conditions were closely similar to those in these preliminary experiments with argon, and in Oliphant's experiments. It is also reasonable to assume that the kinetic energies of the neutral atoms formed by glancing collisions will have values up to that of the incident ions. Further, when the experimental tubes were dismantled, each target was found to have a well defined spot where it had been struck by the beam. In the first tube where the target was not adequately degassed this spot was discoloured, while in later tubes the spot was not discoloured but was pitted. This could hardly have been caused by photons.

In all the experiments a slight instability in the secondary electron current from T to C was observed. This corresponded to fluctuations of the order of 10^{-10} amp. It showed that it would have been pointless to aim at higher sensitivity, which would otherwise have been desirable. It is thought that these fluctuations may have been connected with the fact that it was not possible to degas the inner surface of the canal. To maintain the differential pumping the probe had to be held in a tight fitting glass sheath and therefore could not be treated by the usual methods. The positive ions colliding with the canal walls would cause a localized degassing, but under the circumstances this may well have led to variations in conditions at the canal surface which would produce fluctuations in the intensity of the neutral beam.

§ 3. MAIN OBSERVATIONS

(*a*) *Inert Gases*

The procedure in the main experiments was the same as in those of Oliphant, namely to observe the variation in the electron current between target T and collector C, when the potential difference between them was altered. Retarding curves were obtained in this way for different values of the kinetic energy of the primary beam.

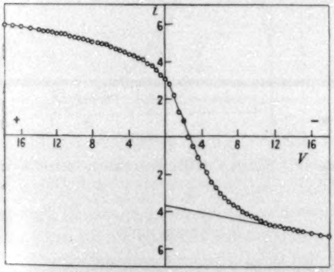


Figure 3. Retarding curve for secondary electron current given by neon atoms. V = potential of collector C with respect to target T; i = secondary electron current, unit 2×10^{-9} amp.

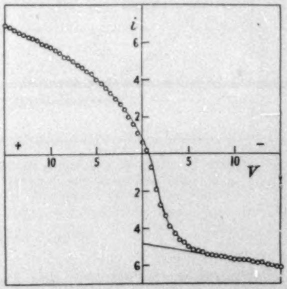


Figure 4. Retarding curve for secondary electron current produced by argon atoms. Unit current $= 2 \times 10^{-9}$ amp.

By comparing the results it was hoped to obtain information on the relative importance of the processes (i) and (ii) discussed in the Introduction. It was found that as the energy of the incident particles was lowered (i.e. as the potential difference applied between A and P (Figure 1) was decreased) the total secondary electron current from T to C decreased. This and the limitation on the sensitivity already discussed (§ 2) set a lower limit to the energy of the beams at which measurements could be made.

The results for typical measurements made with neon, argon and helium are shown in Figures 3, 4 and 5. These are similar to those obtained by

Oliphant (1929) for helium. The fact that the current between the target and collecting cylinder changes sign when the potential difference between them is reversed was interpreted by him as showing that some of the incident metastable atoms are reflected from the target, and cause secondary emission from the surface of the collecting cylinder. The differential curve obtained from the right-hand side of the curve gives the energy distribution of the secondary electrons from the target. Figure 5 for helium shows only this part of the curve. It will be seen that the secondary electron current from the target does not saturate, but becomes a straight line of small constant slope when the voltage is sufficiently increased. This may be due to (1) the unsymmetrical structure of the target–collector system; (2) a slight steady change in the gas pressure in the arc during the time over which a set of observations were being made. The slope of the straight line was different in different runs, and to make the results comparable, the straight line was extrapolated back to zero volts and the retarded electron current taken to be the difference between this extrapolated line and the experimental curve. Because of the unsymmetrical nature of the target–collector system the left-hand side of the curves, which corresponds to an electron current from C to T, does not give much information. Further, some of the incident particles presumably lose kinetic energy on being reflected, and this may also affect the shape of the left-hand side. The lack of symmetry of the target–collector system does, on the other hand, tend to eliminate effects due to multiple reflections between the target and cylinder, so that the ratio of the saturation currents at the beginning of the straight sections is a rough measure of the fraction of the incident metastable atoms reflected from the target. These retarding curves were found to be closely reproducible when the target was degassed by the process described (§ 1). The electron energy distribution curves derived from the retarding curves in Figures 3, 4 and 5 are shown in Figure 6.

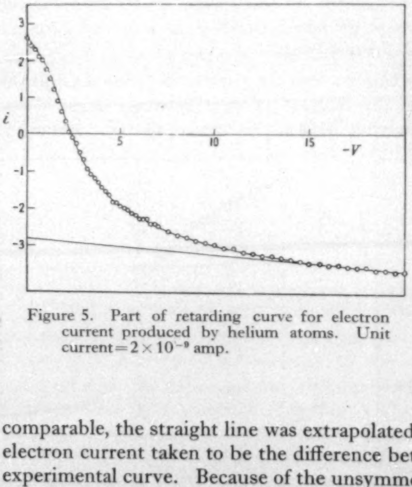


Figure 5. Part of retarding curve for electron current produced by helium atoms. Unit current $= 2 \times 10^{-9}$ amp.

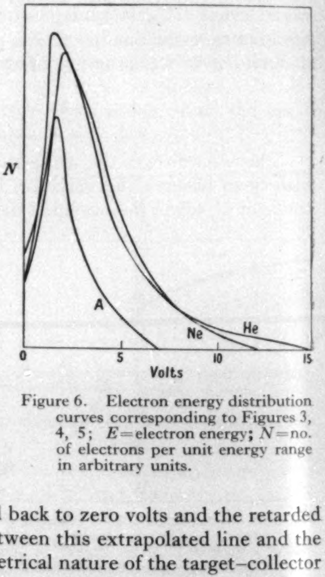


Figure 6. Electron energy distribution curves corresponding to Figures 3, 4, 5; E = electron energy; N = no. of electrons per unit energy range in arbitrary units.

To make the retarding curves obtained using incident particles of various kinetic energies comparable, it is desirable to adjust them so that they have the same maximum ordinate. Families of observations for various energies, adjusted in this way, are shown in Figures 7, 8 and 9.

The energy range which could be covered for helium metastable atoms was rather small, viz. 730 to 1,080 electron volts, chiefly because of the small secondary emission. Within this range the form of the retarding curve was found to be independent of the kinetic energy of the incident particles, as shown by Figure 7.

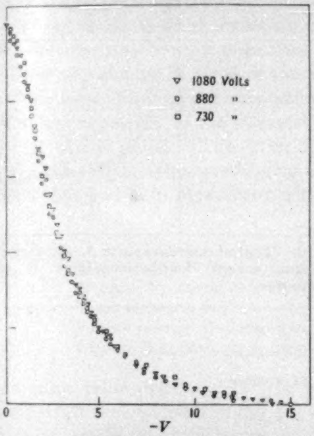


Figure 7. Adjusted secondary electron currents produced by helium.

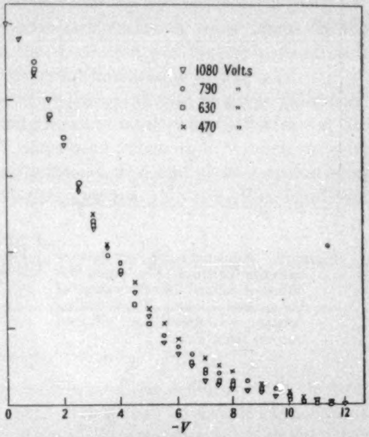


Figure 8. Adjusted secondary electron currents produced by neon.

Figure 8 shows some results for neon, taken over the energy range 470–1,080 electron volts. Here the spread of the experimental points is slightly wider than in Figure 7, but the random placing of the points with respect to the energy of the primary particles seems to show that, within experimental error, the form of the curve is almost independent of this energy.

Figure 9 shows typical results for argon. Here in the range of 0 to 4 collector volts the spread of the experimental points is quite wide, and the points lie in the order of the primary kinetic energies to which they correspond. Above −4 volts the points converge towards a common limit, and their placing becomes more nearly random. Two sets of observations made with incident particles of the maximum and minimum energies used are shown. In both sets of results the differences between corresponding points is smaller than the difference between the corresponding points taken at the next value of the kinetic energy of the primary particles. This shows that the spreading of the experimental points taken at different kinetic energies is outside experimental error.

(*b*) *Hydrogen*

A typical retarding curve obtained using hydrogen is shown in Figure 10. This shows the same general features as were obtained with the inert gases. The corresponding velocity distribution curve is superimposed. A family of retarding

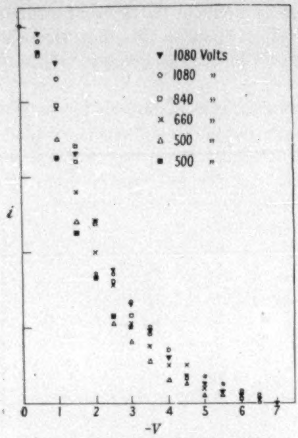


Figure 9. Adjusted secondary electron currents produced by argon for different values of kinetic energy of the primary beam. V = retarding voltage; i = secondary electron current from T.

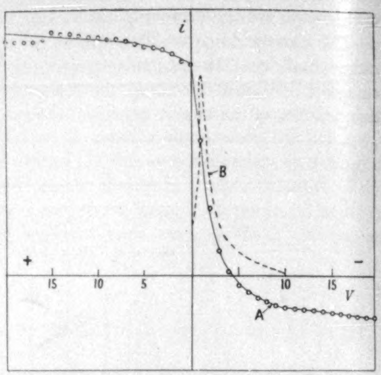


Figure 10. Typical retarding curve A, and derived electron energy distribution curve B for hydrogen.

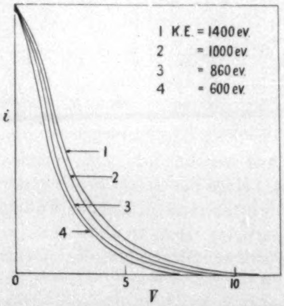


Figure 11 Adjusted secondary electron currents produced by hydrogen atoms and molecules of various kinetic energy.

curves for hydrogen, adjusted to a common maximum ordinate, is shown in Figure 11. The experimental points lie closely on the curves. The upper limit of the energy of the secondary electrons increases with the kinetic energy of the primary particles, and the experimental points follow the order of the primary particle energy, for all values of the collecting field.

§4. DISCUSSION OF RESULTS

(*a*) *Inert Gases*

The maximum energy a metastable atom can give to an electron in the metal, apart from any contribution due to its kinetic energy, is given by the energy of the metastable state less the work function of the metal. These values for helium, neon and argon are compared with the observed upper energy limits of the secondary electrons in Table 1. The accuracy of the observed values for the upper energy limits is limited to the nearest 0·5 volt by two factors: (i) the difference between successive galvanometer readings at 0·5 volt intervals at this part of the curve is approaching the least measurable change in current; (ii) it is difficult to decide exactly at what point on the curve saturation is reached, because the approach to saturation seems to include a small asymptotic contribution. With these limitations, the calculated and observed upper energy limits are in good agreement.

There is no definite evidence for secondary emission from atoms in the 20·5-volt metastable state of helium but the resolution was probably insufficient to show any change in slope of the experimental line arising in this way.

The lower limit of the secondary electron energy should be given (Oliphant 1929) by the energy of the metastable state less the total potential barrier at the surface of the metal. The latter has a calculated value of 17·9 electron volts (Oliphant 1929); this would give a minimum electron energy of zero for neon and argon, and is in agreement with the observations. On the other hand, the minimum value predicted for the electrons liberated by helium is 1·9 electron volts, while the observed minimum was zero. There may be several explanations for this: (i) the potential barrier at the surface of the metal may not have the value assumed; (ii) if the electrons were scattered in emerging from the surface, this might reduce their energy; (iii) secondary electrons liberated by the kinetic energy of the incident particles would have zero minimum energy, and the presence of these might mask the lower energy limit for individual interactions; (iv) a contact potential between the target and collecting cylinder might shift the voltage scale enough to mask the presence of a non-zero minimum energy; it is unlikely that this could account completely for the failure to observe the predicted minimum energy, because the presence of a contact potential difference of the magnitude required would also produce an appreciable alteration in the value of the observed upper energy limit.

Table 1

Gas	He	Ne	A
Energy of metastable state (ev.)	19·8	16·5	11·5
Work function (v.)	4·4	4·4	4·4
Excess energy available (ev.)	15·4	12·1	7·1
Observed upper energy limit (ev.)	15·0	12·0	7·0

The consistency of the results obtained using helium metastable atoms of various kinetic energies would seem to show that the kinetic energy of the particles made no appreciable contribution to the observed secondary electron emission. The form of the observed energy distribution curve is nearer to that obtained by Oliphant than to that calculated by Massey. The results for neon are also almost independent of the kinetic energy of the incident metastable atoms. The results for argon show a definite tendency for the proportion of the higher energy electrons in the region 0–4 electron volts to increase with increase in kinetic energy of the bombarding particles. For still higher energies of the secondary electrons, the energy distribution curves converge to a common upper limit. It therefore seems likely that the kinetic energy of the argon metastable atoms does contribute to the observed secondary emission by the process (ii) discussed in the Introduction. We would expect the mean energy of the electrons liberated in this way to increase with increase in the kinetic energy of the incident particles, and this also is in agreement with the observed results. Apparently these 'thermionic' electrons have not enough energy to change appreciably the high energy tail of the energy distribution. The difference between the results for argon and those for neon and helium could be due either to the greater mass of the argon atom or to the smaller excess energy available for individual interactions between argon metastable atoms and metallic electrons.

The ratio of the saturation current from C to the sum of the negative and positive saturation currents from C and T gives an approximate measure (§ 2) of the fraction of the metastable atoms reflected from the target. Its magnitude for different values of the primary kinetic energy is shown in Figure 12. For all the gases the proportion of the incident particles reflected from the target decreases with increase in their kinetic energy. This effect is most marked for argon, and may be connected with the greater variation in the form of the secondary electron energy distribution.

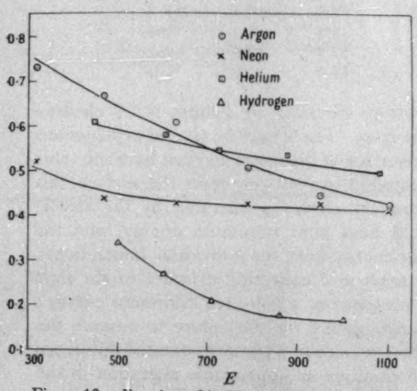


Figure 12. Fraction of incident particles reflected from target as a function of incident kinetic energy, *E* volts.

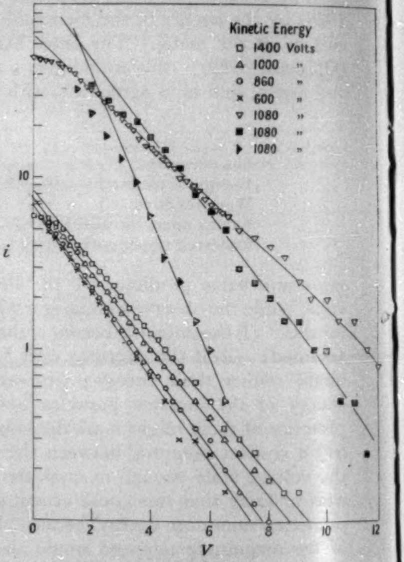


Figure 13. Semi-logarithmic plots of retarded electron currents i against retarding voltage V. The top four symbols refer to hydrogen, the remainder to helium, neon and argon in descending order.

(*b*) *Hydrogen*

The beam of neutral particles formed in the present apparatus almost certainly contained both atoms and molecules, and investigations concerned with the design of proton sources (Lamar, Samson and Compton 1935) show that it might be predominantly molecular. The results obtained with hydrogen are different from those given by the inert gases, in that the whole of the retarding curve shifts with the kinetic energy of the incident particles. Semi-logarithmic plots of the retarding curves were found to give good straight lines. Some of these are shown in Figure 13. For comparison semi-logarithmic plots of some retarding curves taken with the inert gases are also shown. Those for argon and neon show a greater departure from linearity than those for hydrogen and helium.

The fact that the semi-logarithmic plots for hydrogen are linear over nearly the complete range of voltages may not be very significant because any rapidly falling retarding curve would tend to give such a result. Assuming, however, that the electron emission is of the second type discussed in the Introduction the electron temperature can be determined from the slope of the semi-logarithmic plot, and this in turn can be taken to be the effective temperature of the emitting surfaces. This description of the results is only an approximation as an energy distribution with a finite upper limit is not truly Maxwellian.

The results obtained for hydrogen are summarized in Table 2. These show that the electron temperature increases with the energy of the incident particles, though not very uniformly, and seems to be approaching a limiting value. The upper energy limits also seem to approach a limiting value of about 11 or 12 volts.

Table 2

Max. energy of incident particles (ev.)	Electron temp. (° K.)	Mean electron energy (ev.)	Mean electron temp. (° K.)	Max. electron energy (ev.)
600	20,600	2·7	20,600	7·5
710	21,500	2·8	21,500	8
860	21,500	2·8	21,500	8·5
1000	24,800	3·2	25,100	9
1020	23,800	3·1		10
1020	25,400	3·3		10
1000	26,700	3·4		10
1150	25,000	3·25	25,000	11
1350	24,000	3·1	25,100	12
1400	27,800	3·6		11
1400	24,700	3·2		11

It can be seen from Figure 13 that all the semi-logarithmic plots depart from linearity in the neighbourhood of zero collecting volts, and that this effect becomes more pronounced as the energy of the incident particles is increased. Although the zero point in these curves, as for the inert gases, is a little indefinite, we would not expect this to vary with the kinetic energy of the primary particles, where the total power delivered by the beam is small; it therefore seems likely that the effect is real. This suggests that more than one process is involved in the production of the secondary electrons.

So far as the author knows, the possibility that direct interactions between ordinary excited atoms and metallic electrons could produce secondary emission in the present type of tube has not previously been considered. If the time required for the primary particles to travel from the canal (where they are neutralized) to the target were of the same order of magnitude as the mean life-time of an ordinary excited state, and if excited atoms or molecules were produced by neutralization, such interactions would be possible and could cause secondary electron emission if the energy of excitation were greater than the work function of the metal. Calculated values of the mean life-times of the excited states of hydrogen atoms are given by Condon and Shortley (1935). The mean life-times increase with the degree of excitation, the value given for $n=6$, for example, being $1{\cdot}96\times10^{-7}$ second. This value is obtained by assuming that equal numbers of atoms exist in all the states corresponding to $n=6$. The life-times of the individual states may be considerably greater than this value, for example, the value given for the 6h state is $6{\cdot}1\times10^{-7}$ second. The transit times for 1,000 and 1,400 electron volt hydrogen atoms travelling between the canal and target (a distance of 10 cm.) are $2{\cdot}88\times10^{-7}$ and $1{\cdot}92\times10^{-7}$ second respectively. It would, therefore, be possible for some hydrogen atoms in the higher excited states to reach the target before decay. The excitation potentials of these states go up to 13·54 volts, the ionization potential, and since this is well above the work function of molybdenum it seems that they could produce secondary emission by interactions of type (i). Photons produced by the decay of excited atoms in transit between the canal and surface could also produce electron emission. Only a small proportion of these photons would, however, reach the target. Secondary electrons produced by direct interactions with excited atoms might thus be responsible for the non-Maxwellian contribution to the total emission for the greatest secondary electron energies.

The maximum energy an excited hydrogen atom could give a secondary electron is the ionization energy less the work function of the target, i.e. 9·1 electron volts. This could not account for the upper energy limit of 11–12 electron volts to which the results tend. The ionization potential of the hydrogen molecule is 15·6 volts, so that the maximum energy an excited hydrogen molecule could give to a metallic electron is 11·2 electron volts. This agrees well with the observed result. The mean life-times of the excited states of the hydrogen molecule are not known, but if, as is likely, they are of the same order of magnitude as those of the atom, secondary emission by individual interactions would be possible.

Since the transit times for helium atoms are twice those for hydrogen atoms, it seems possible that the ordinary excited states might cause secondary emission for the higher kinetic energies of the bombarding particles. There is, however, no definite evidence for this from the present work. This might be due to a tendency of the neutralization process to leave the helium atoms in the metastable state. Neon and argon would probably give transit times which are too long to allow any effects due to excited atoms.

A further mechanism that has to be considered for the ejection of secondary electrons is that in which elastic collisions between incident atoms and metallic electrons give the latter sufficient energy to penetrate the potential barrier after reversing their direction of motion in an elastic collision. The maximum energy a hydrogen atom with kinetic energy of 1,000 electron volts could give an electron in an elastic collision is about two electron volts; since this is less than the work function of the metal, it would seem that this mechanism can be ruled out.

The retarding runs taken with hydrogen give similar evidence to that obtained with the inert gases for reflection of the incident particles from the target. The reflection coefficients determined as before are shown in Figure 12. The existence of an electron current from the collector to the target in this case might also be due in part to photons produced by excited atoms in the beam decaying inside the collecting cylinder.

§5. SUMMARY

The results reported here generally confirm those of Oliphant and of Chaudhri and Khan. The electron energy distribution given by helium metastable atoms agrees with that observed by Oliphant rather than that calculated by Massey. The present results for hydrogen are comparable with those of Chaudhri and Khan for potassium, because neither atom has a metastable state, and both agree in showing the secondary electrons to have a nearly Maxwellian energy distribution. The results for neon are similar to those for helium in that the form of the electron energy distribution does not vary appreciably with the energy of the primary particles, while the proportion of high energy electrons liberated by argon metastable atoms increases with their kinetic energy. The similarity of the results for hydrogen to those of Chaudhri and Khan (1948) for mercury is of particular interest, because while it is known that metastable atoms of thermal energy can liberate secondary electrons (Webb 1924), the emission produced by high-speed mercury metastable atoms seems to be due almost entirely to their kinetic energy.

The general conclusion that can be drawn from the data at present available seems to be that the relative efficiency of the individual interaction process, and the mechanical transfer process, varies with the excess energy available for the former. For neon and helium the contribution made by the second process is negligible; for argon there is a small proportion of the secondary emission caused by the second process; for mercury, where the excess potential energy available has a slight positive value the emission seems to be almost entirely due to the quasi-mechanical transfer process, while hydrogen and potassium give results which are also explicable in terms of the latter process. There is also evidence that hydrogen atoms and molecules can produce secondary electrons by individual interactions, where ordinary excited atoms can reach the target. These conclusions again are in agreement with work on the secondary electron emission produced by positive ions (Oliphant and Moon 1930), where the relative importance of the emission produced by individual interactions and by the kinetic energy of the incident particles varies in the same way.

ACKNOWLEDGMENTS

The author wishes to thank Professor K. G. Emeléus for suggesting and supervising this work, Dr. R. H. Sloane for advice on many points, and Professor H. S. W. Massey for a stimulating discussion. The work was done whilst the author was a Senior Research Student of Queen's University. Part of the cost of batteries was defrayed from the Research Fund of the University.

REFERENCES

Chaudhri, R. M., and Khan, A. W., 1948, *Proc. Phys. Soc.*, **61**, 526.
Condon, E. U., and Shortley, G. H., 1935, *The Theory of Atomic Spectra* (Cambridge: University Press), p. 346.
Lamar, E. S., Samson, E. W., and Compton, K. T., 1935, *Phys. Rev.*, **48**, 886.
Massey, H. S. W., 1930 *Proc. Camb. Phil. Soc.*, **26**, 386; 1931, *Ibid.*, **27**, 460.
Oliphant, M. L. E., 1929, *Proc. Roy. Soc.* A, **124**, 228.
Oliphant, M. L. E., and Moon, P. B., 1930, *Proc. Roy. Scc.* A, **127**, 388.
Sloan, D. H., Thornton, R. L., and Jenkins, F. A., 1935, *Rev. Sci. Instrum.*, **6**, 75.
Uyterhoeven, W., and Harrington, M. C., 1930, *Phys. Rev.*, **35**, 438, and **36**, 709.
Webb, H. W., 1924, *Phys. Rev.*, **24**, 113.

The Variation with Temperature of the Piezoelectric Coefficients of Quartz

By A. C. LYNCH
Post Office Engineering Research Station, London N.W.2

MS. received 14*th June* 1950

ABSTRACT. The equivalent electrical circuits of three bars in longitudinal vibration were measured at approximately 25, 52·5 and 80° c. In this range of temperature the temperature coefficients of d_{11} and d_{14} are respectively -2_{30} and $+13_{60}$ parts/million/° c., and there is no evidence to support Cady's suggestion that d_{11} passes through a maximum near room temperature.

The measurements suggest a rather high value for d_{14}: $(-2{\cdot}21 \pm 0{\cdot}1) \times 10^{-8}$ cm/E.S.U. of potential at 25° c.

---

§ 1. INTRODUCTION

Widely varying results have been reported for the piezoelectric coefficient d_{11} of quartz at high and low temperatures, while no values have been published for the variation with temperature of d_{14} (Cady 1946, p. 221). A few results, expressed as temperature coefficients, are :

Observers	Date	Range of temperatures (°c.)	Temperature coefficient of d_{11} (parts/million/°c.)
Andreeff, Fréedericksz, and Kazarnowsky	1929	15 to 500	−350
Van Dyke	1932	−80 to 40	Small at −80°c., −1000 at 40°c.
Langevin	1936	20 to 200	−550
Langevin and Moulin	1937	−60 to 23	Zero between 0 and 20°c., +1000 below 0°c.
Clay and Karper	1937	17 to 90	−110

Cady quotes the last of these as −7, owing to a misreading of Clay and Karper's notation. He concludes, from these and other reported results, that the coefficient probably passes through a very flat maximum near room temperature. Since there is no crystallographic transition to cause a discontinuity in the variation, we might expect a roughly parabolic curve for piezoelectric coefficient against temperature. If so, measurements at three temperatures should show whether a maximum exists, not only if it lies between the temperature of measurement, but also if it is somewhat outside them. Apparatus was available with which the equivalent electrical circuit of quartz resonators could be measured with sufficient accuracy in the temperature range 25 to 80°c.

In these measurements we need not attempt to find the absolute value of the piezoelectric coefficient, but only its variation with temperature. Hence we need not consider the approximations made in deducing it from the equivalent circuit, nor any mechanical effects introduced by the electrodes, and the specimens need not be those best suited to the measurement of the piezoelectric coefficient itself. If the measurements lead to a value for d_{11} agreeing with the accepted figure of $6{\cdot}9 \times 10^{-8}$ cm/E.S.U. of potential (Cady 1946, p. 219)—as in fact they did—then we know the approximations are sufficiently good at 20° c.; we are then assuming only that they are equally justified at other temperatures. If the experiments had disclosed a non-linear variation of d with temperature, this assumption would need further examination; but as d was found to vary linearly, there is no reason to doubt the assumption.

§2. EXPERIMENTS

For the sake of simplicity in interpreting the results, three bars were chosen for test, one having its length along the conventional Y-axis, and the others at 45° and at 135° to the Y-axis in the YZ plane. If an alternating electric field is applied along the X-axis, the bars can be set into longitudinal vibration; the behaviour of the Y-axis bar then gives a measurement of the coefficient d_{12}, which in quartz is equal to $-d_{11}$, while the results for the other two bars, taken together, give values for both d_{11} and d_{14}.*

Table 1 gives their dimensions and frequencies of resonance.

Table 1

Direction of length of bar	Dimensions (cm.)	Approximate resonance frequency (kc/s.)
Y-axis	$3{\cdot}622\times0{\cdot}503_5\times0{\cdot}187_8$	75
45° to Y-axis	$2{\cdot}001\times0{\cdot}300_5\times0{\cdot}204_3$	139
135° to Y-axis	$2{\cdot}001\times0{\cdot}300_5\times0{\cdot}204_0$	174

Silver was evaporated on to the faces in the YZ plane to serve as electrodes and the bars were held between domed silver contacts mounted on light springs. Thus mounted, the bars were set up in a lagged copper box, heated by a current of warm air. Temperatures were measured by a mercury thermometer mounted with its bulb close to the specimen, and could be kept within about 0·2 deg. for periods of a quarter of an hour. The specimens took up each new temperature within a few minutes (as judged by observations of their frequencies of series resonance), but were left at approximately the required temperature for about two hours before measurements were begun.

Measurements of equivalent capacitance at frequencies near that of series resonance were made as described by Lynch (1950). The only difficulty was that the oblique bars had large temperature coefficients of frequency, and special care was needed to keep their temperatures constant and to work quickly. Closely reproducible results were eventually obtained, however, and are given in Table 2.

The capacitances have been corrected, by interpolation, to values for three uniformly spaced temperatures. C is the measured capacitance in the equivalent circuit in pF.; f_0 is the frequency of series resonance in kc/s.; d' is the piezoelectric coefficient deduced†, with positive or negative sign chosen to make d_{11} positive; d_{14} and the second value for d_{11} are obtained from d' for the oblique bars; d', d_{11}, and d_{14} are in 10^{-8} cm/E.S.U. of potential. The density of quartz has been taken as $2{\cdot}64_9$ at 20° C. (Cady 1946, p. 411), which is approximately

---

* Cady (1946, p. 212) gives, for a bar rotated through an angle θ round the X-axis, $d_{12}' = -c^2 d_{11} + scd_{14}$ where c, $s = \cos\theta$, $\sin\theta$. Hence, for the two specimens mentioned, $d_{11} = -[(d')_{45} + (d')_{135}]$ and $d_{14} = (d')_{45} - (d')_{135}$ where $(d')_{45}$ and $(d')_{135}$ are the piezoelectric coefficients for the 45° and 135° bars.

† From equations given by Cady, 1946, pp. 88 and 297, we have

$$d' = \pm\left(\frac{\pi^2 Ct}{32l^3 bf_0^2 \rho}\right)^{1/2}$$

where l, b, and t are the length, breadth, and thickness of the bar, ρ its density, and f_0 its frequency in series resonance. C must now be in E.S.U.

the temperature at which the dimensions were measured. At other temperatures, both the density and dimensions are different, but the joint effect of their changes on the capacitance C' is zero for the Y-axis bar, and only 6 parts/million/° c. for the other two bars.

Table 2

Temperature (° c.)		25		52·5		80
Y-axis bar :	C	$0{\cdot}0327_8$		$0{\cdot}0323_7$		$0{\cdot}319_6$
	f_0	$74{\cdot}85_1$		$74{\cdot}85_0$		$74{\cdot}84_1$
	$d_{11}=-d'$	$6{\cdot}94_5$		$6{\cdot}90_2$		$6{\cdot}85_9$
Oblique bars : 45° :	C	$0{\cdot}00453_1$		$0{\cdot}00455_6$		$0{\cdot}00457_0$
	f_0	$138{\cdot}59$		$138{\cdot}31$		$138{\cdot}00$
	d'	$-4{\cdot}58_0$		$-4{\cdot}60_2$		$-4{\cdot}62_1$
135° :	C	$0{\cdot}00192_4$		$0{\cdot}00181_0$		$0{\cdot}00170_0$
	f_0	$174{\cdot}16$		$173{\cdot}76$		$173{\cdot}34$
	d'	$-2{\cdot}37_4$		$-2{\cdot}30_9$		$-2{\cdot}24_2$
Hence	d_{11}	$6{\cdot}95_4$		$6{\cdot}91_1$		$6{\cdot}86_3$
	d_{14}	$-2{\cdot}20_6$		$-2{\cdot}29_3$		$-2{\cdot}37_9$
Differences for 27·5°c. :						
	d_{11}		$-0{\cdot}04_3$ $-0{\cdot}04_3$		$-0{\cdot}04_3$ $-0{\cdot}04_8$	
	d_{14}		$-0{\cdot}08_7$		$-0{\cdot}08_4$	

Hence the mean temperature coefficients of d_{11} and d_{14} are respectively -2_{30} and $+16_{60}$ parts/million/° c. in the range 25 to 80° c.

These results suggest that any maximum in d_{11} is at a temperature well outside the range 25 to 80° c. (probably at least 100 deg. away).

Although these experiments were not intended as a measurement of the piezoelectric coefficients, the oblique bars give a useful indication of d_{14}; for, provided only that the orientations of the bars are as stated, any error in d_{14} would occur equally in d_{11}; so that, since d_{11} determined from these bars agrees with the figure for the Y-axis bar and also with the accepted figure, the value for d_{14} should be equally reliable. The value $(-2{\cdot}21 \pm 0{\cdot}1) \times 10^{-8}$ cm/E.S.U. of potential at 25° c., is higher than most other reported values, and higher than the value $-2{\cdot}0$ recommended by Cady (1946, p. 219), though it is lower than the value $-2{\cdot}56$ found by Mason (1943).

§ 3. ACKNOWLEDGMENT

Acknowledgment is made to the Engineer-in-Chief of the General Post Office for permission to make use of the information contained in this paper.

REFERENCES

Andreeff, A., Fréedericksz, V., and Kazarnowsky, I., 1929, *Z. Phys.*, **54,** 477.

Cady, W. G., 1946, *Piezoelectricity* (New York : McGraw-Hill).

Clay, J., and Karper, J. G., 1937, *Physica*, **4,** 311.

Langevin, A., 1936, *J. Phys. Radium*, sér. 7, **7,** 95.

Langevin, A., and Moulin, A. M., 1937, *J. Phys. Radium*, sér. 7, **8,** 257.

Lynch, A. C., 1950, *Proc. Phys. Soc.* B, **63,** 323.

Mason, W. P., 1943, *Bell Syst. Tech. J.*, **22,** 178.

Van Dyke, K. S., 1932, *Phys. Rev.*, **42,** 587.

Solvent Effects in Dipole Moment Measurements

By I. G. ROSS* and R. A. SACK†

* Department of Chemistry, University of Sydney; now at Department of Chemistry, University College, London.

† Division of Electrotechnology, Commonwealth Scientific and Industrial Research Organization, University Grounds, Chippendale, N.S.W.

Communicated by W. E. Duncanson; MS. received 9th May 1950

ABSTRACT. Previous theories advanced to account for the differences between the moments of polar molecules, as measured in dilute solution (in non-polar solvents) and from gas measurements, are discussed on the basis of an internal field function ξ, determining the field inside an ellipsoidal obstacle in a polarized uniform medium. Contours of ξ as a function of the axial ratios of the ellipsoid have been calculated. Onsager's theory of the static dielectric constant is extended to account for the solvent effect in ellipsoidal solute molecules of uniform polarizability, in which the dipole is parallel to one of the axes of the ellipsoid; the new equation thus obtained gives results in better agreement with experimental data than previous theoretical treatments.

The rule that the sign of the solvent effect of a substance should be opposite to that of its Kerr constant is shown not to be of general applicability.

§ 1. INTRODUCTION

A THEORY of the dielectric constant of vapours, liquids and solutions, the molecules of which have permanent electric moments, has been developed by Debye (1929). Although Debye's theory in the case of liquids leads to conclusions which are in qualitative disagreement with experimental facts and has therefore been largely superseded by the later theory of Onsager (1936), it is still usually employed in the determination of dipole moments from measurements on gases or on dilute solutions in non-polar media. For most substances, however, the application of this theory leads to a slight discrepancy between the values of the dipole moments μ_{gas} and μ_{soln} obtained by the two experimental methods. The difference $\Delta\mu = \mu_{\text{soln}} - \mu_{\text{gas}}$, which may be called the solvent effect, is generally negative, but for a minority of substances positive values are found; in either case the magnitude of the solvent effect is dependent on the dielectric constant of the solvent.

It is generally recognized that any theory of the solvent effect must take into account, not only the dielectric constant of the medium, but also particularly the structure of the solute molecules. In theoretical treatments where this factor is taken into consideration a molecule is frequently treated as an ellipsoidal solid with dimensions most closely approximating its actual shape. Such a model is mathematically tractable, and has the further advantage that when placed in a uniform electric field, the field within it is also uniform. Consequently the higher moments in the multipole development of the charge distribution within the molecule are immaterial. The relation between the field inside the ellipsoid and the polarization of the surrounding medium is determined by certain coefficients, hereafter referred to as internal field functions, which depend on the shape of the molecule and its orientation with regard to the field; they are defined in § 2 and evaluated in the Appendix. These internal field functions arise irrespective of whether the problem of the solvent effect is approached from the point of view of Debye's or of Onsager's theory. The former approach has been used by previous authors; their procedure is critically discussed in §3. It appears more logical, however, to extend Onsager's theory to the case of dilute solutions; this is done in §4, and the results obtained are discussed and compared with experiment in §5. Finally in §6 a comparison is made between the formulae for the values of the solvent effect and the Kerr constant which shows that the two are not directly related, in contrast to previous statements.

§2. THE INTERNAL FIELD FUNCTIONS ξ.

As shown by Maxwell (1873), the disturbance of electric potential produced in a uniform electric field by a small ellipsoidal obstacle of uniform polarizability and with semi-axes *a*, *b*, *c*, depends on the quantities

$$\left.\begin{aligned}\xi_a&=\tfrac{1}{2}abc\int_0^\infty\frac{d\lambda}{(a^2+\lambda)^{3/2}(b^2+\lambda)^{1/2}(c^2+\lambda)^{1/2}},\\ \xi_b&=\tfrac{1}{2}abc\int_0^\infty\frac{d\lambda}{(a^2+\lambda)^{1/2}(b^2+\lambda)^{3/2}(c^2+\lambda)^{1/2}},\\ \xi_c&=\tfrac{1}{2}abc\int_0^\infty\frac{d\lambda}{(a^2+\lambda)^{1/2}(b^2+\lambda)^{1/2}(c^2+\lambda)^{3/2}}.\end{aligned}\right\}\quad\ldots\ldots(1)$$

In the case where we are dealing with an ellipsoidal *cavity* in a uniformly polarized medium, ξ has the simple meaning that the internal field within the cavity is

$$F=E+4\pi\xi P,\qquad\ldots\ldots(2)$$

where $\xi=\xi_a$, ξ_b, or ξ_c according as the field E is applied parallel to the axis *a*, *b*, or *c*, and P is the polarization. For this reason we may call ξ the ***internal field function.***

For arbitrary values of the axes the following identity holds:

$$\xi_a+\xi_b+\xi_c=1;\qquad\ldots\ldots(3)$$

hence for spherical obstacles all these three quantities (1) become equal to $\frac{1}{3}$ (cf. the Lorentz–Lorenz internal field for a spherical cavity, $F=E+4\pi P/3$).

Explicit expressions for ξ have been given to date only for the cases where two axes are equal, i.e. for spheroidal obstacles (Maxwell 1873—see Appendix, equations (60), (61)). For unequal values of *a*, *b*, *c*, the formulae (1), despite their simple form, are not suitable for computing ξ as tables of the integrals are not available. General explicit expressions for ξ are derived in the Appendix (equations (46), (54), (58)).

Values of ξ, as a function of the axial ratios, have been evaluated with the use of these formulae, and are shown in Figure 1. In this figure A is the length of the axis parallel to which the field is applied, and B and C are the greater and smaller, respectively, of the other two axes; the dotted lines represent the various limiting spheroidal cases, points on which are evaluated by means of (60) and (61). Values for $A/B>1$ are given for a short range only, since values in this region may be calculated readily from the identity (3).

§3. REVIEW OF PREVIOUS THEORIES OF THE SOLVENT EFFECT

Debye's theory, which assumes that the polar molecules are oriented under the influence of the Lorentz field $F=E+4\pi P/3$, leads to the well-known formula for the dielectric constant ϵ_{soln} of a liquid mixture

$$\frac{\epsilon_{\text{soln}}-1}{\epsilon_{\text{soln}}+2}=\frac{4\pi}{3}\Sigma N\left(\alpha+\frac{\mu^2}{3kT}\right),\qquad\ldots\ldots(4)$$

where α and μ denote the polarizability and the permanent dipole moment, respectively, of the individual molecules, N their number per unit volume, and the summation extends over all types of molecules present (Debye 1929). To account for the solvent effect, i.e. the difference between the values of μ calculated according to (4) from measurements on dilute solutions and on vapours, a number of empirical relations have been derived, the most satisfactory of which is that due to Goss (1937, 1940); the latter has some theoretical justification, and also takes into account the anisotropy of the medium. In addition the problem has been approached theoretically from two viewpoints.

The first approach is based on the theory of Raman and Krishnan (1928) concerning the refractive index and dielectric constant of polar liquids. This theory considers in detail the internal field acting on each dipole and its deviation from the Lorentz field. It has been adapted to mixtures, especially of polar substances with non-polar liquids, by Goss (1934) and other workers. In particular, if small correction terms arising from the anisotropies of the solvent and solute are neglected, then at infinite dilution there is obtained the relation (Le Fèvre and Le Fèvre 1935)

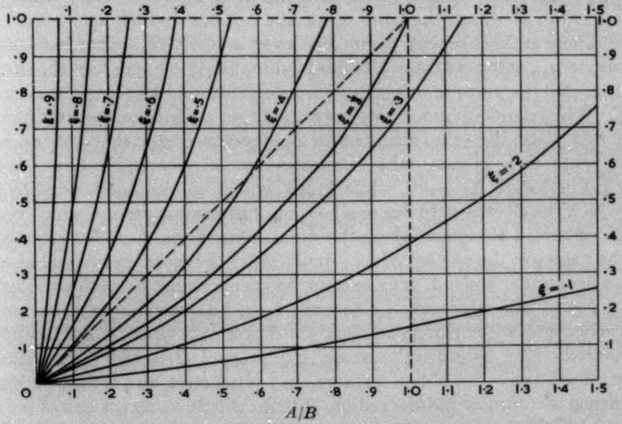


Figure 1. Contours of the internal field function.

$$ {}_0P_{\text{soln}}/{}_0P_{\text{gas}} = 1 + 3\Theta(\epsilon-1)/4\pi\mu^2(\epsilon+2), \qquad \ldots\ldots(5)$$

where ${}_0P$ is the orientation polarization of the solute, μ its dipole moment, ϵ the dielectric constant of the solvent, and

$$\Theta = \sum_i \mu_i^2\sigma_i + \sum_{i\neq j} \mu_i\mu_j q_{ij} \quad i, j = 1, 2, 3.$$

Here the μ_i are the components of $\boldsymbol{\mu}$ along the principal axes of the ellipsoid representing the molecule, q_{ij} are the components of the polarization field tensor, and $\sigma_i = q_{ii} - 4\pi/3$, i.e. $\sigma_i P$ is the excess of the internal field over the Lorentz–Lorenz internal field for a spherical cavity. For the case where the dipole is parallel to one of the principal axes of the ellipsoid—and for simplicity we shall confine the discussion to such molecules—and the solvent medium is isotropic, $\Theta=\mu^2\sigma_i=4\pi\mu^2(\xi_i-\frac{1}{3})$, and the expression (5) reduces to

$${}_0P_{\text{soln}}/{}_0P_{\text{gas}}=1+3(\xi-\tfrac{1}{3})(\epsilon-1)/(\epsilon+2)$$

or, since ${}_0P$ is proportional to μ^2,

$$S_{\text{RK}}=\frac{\mu_{\text{soln}}}{\mu_{\text{gas}}}=1+\tfrac{3}{2}(\xi+\tfrac{1}{3})\frac{\epsilon-1}{\epsilon+2}+\dots \qquad \dots\dots(6)$$

The Raman–Krishnan theory, both in the simplified form (5) and in the exact form, in which allowance is made for anisotropy effects, has been critically tested by Jenkins and Bauer (1936), who find that the observed values of Θ, as calculated from experimental values of ${}_0P_{\text{soln}}$ in six non-polar solvents and ${}_0P_{\text{gas}}$ may vary within the range of ± 50% of the mean value.

An alternative theory, based on a consideration of the moment induced in the solvent by the dissolved polar molecule, has been developed by Frank (1935) and Higasi (1936), of whom the latter especially has been widely quoted in the literature. These authors replace the molecules by a cavity of appropriate shape, with a dipole $\mu=\mu_{\text{gas}}$ at the centre. The field due to this dipole polarizes the surrounding medium, and the volume integral of this polarization, $\Delta\mu$, is taken as the solvent effect. The results obtained by the two authors are inconsistent for two reasons. In the first place, the value obtained for $\Delta\mu$ depends on the shape of the volume of the solvent; thus Higasi, who integrates over a spherical volume, finds, for a spherical molecule, $\Delta\mu/\mu=0$, while Frank, whose volume is bounded by a surface over which the electric field due to μ has constant magnitude and which resembles a prolate spheroid elongated along the dipole axis, finds $\Delta\mu/\mu=5\%$ (in benzene); and the values of the solvent effect calculated by the two theories will always differ by this amount. Indeed, even greater differences may be predicted by the consistent use of this approach. The ratio $\Delta\mu/\mu$ becomes larger as the enveloping surface becomes more elongated along the dipole axis, until, in the limiting case of an infinitely long cylindrical surface $\Delta\mu/\mu=(\epsilon-1)/3\epsilon=18{\cdot}6\%$ (in benzene) for spherical solutes. For oblate surfaces, $\Delta\mu/\mu$ is negative, and in the limiting case of a cylindrical medium of infinite radius, with the dipole along the axis of the cylinder, $\Delta\mu/\mu=-37{\cdot}2\%$.

The reason for this is that even at large distances the contribution to $\int \mathbf{P}\,dv$ due to that part of the solvent beyond the chosen boundary surface does not tend to zero. If the volume is assumed spherical and the distribution of polarizable matter isotropic as in the original Clausius–Mosotti–Debye theory, induced polarization does not come into play. The correct approach, on this treatment, is therefore to find the apparent dipole moment *on the Clausius–Mosotti model*, which is that of a spherical unit, composed in this case of the solute molecule, plus the additional solvent contained between it and an enclosing spherical surface (whose radius is immaterial). Thus it appears that Higasi has chosen the correct boundary surface, although, in assuming that $\int \mathbf{P}\,dv$, over the volume between his spherical outer boundary 'at infinity' and the actual boundary of the liquid, is zero, his reason for so choosing is incorrect.

Higasi, however, is in error in evaluating $\Delta\mu$. If $\mathbf{E}$ is the field due to $\boldsymbol{\mu}$ at a given point, then, following Frank,

$$\mathbf{P}=(\epsilon-1)\,\mathbf{E}/4\pi \quad \text{and} \quad \Delta\boldsymbol{\mu}=(\epsilon-1)\int \mathbf{E}\,dv/4\pi.$$

According to Higasi, however, $\mathbf{P}=N\alpha\mathbf{E}=(\epsilon-1)\,\mathbf{E}/(\epsilon+2)$ where N is the number of molecules per unit volume and α their polarizability. But the polarizing field at a point should be the spherical cavity field $\mathbf{F} = \mathbf{E} + 4\pi\mathbf{P}/3$; moreover, in evaluating $\mathbf{E}$, Higasi neglects the influence of the dielectric constant of the intervening medium, and his value for $\Delta\boldsymbol{\mu}$ should be further divided by ϵ on this account. The combined effect of these two factors is to multiply the Higasi expression by $(\epsilon+2)/3\epsilon$, which leads to the same result as the Frank calculation applied to a spherical volume. Inspection of Higasi's integrals for ellipsoidal cavities shows that its equation can be simply written

$$\mu_{\text{soln}}/\mu_{\text{gas}} = 1 + 3\left(\xi - \tfrac{1}{3}\right)(\epsilon - 1)/(\epsilon + 2), \qquad \ldots\ldots(7)$$

or in the amended form,

$$S_{\mathrm{H}} = \mu_{\text{soln}}/\mu_{\text{gas}} = 1 + \left(\xi - \tfrac{1}{3}\right)(\epsilon - 1)/\epsilon. \qquad \ldots\ldots(8)$$

As the theory developed below, in which the Clausius–Mosotti approach is not used, leads under certain conditions to formula (7), it is appropriate to draw attention here to the similarity between the Raman–Krishnan (6) and Higasi (7) equations, the latter of which predicts values of $\Delta\mu$ exactly twice those given by the former.

§4. A NEW THEORY OF THE SOLVENT EFFECT

Onsager (1936), in his paper on the static dielectric constant of pure liquids and solutions, in which he considers the polar molecules as being oriented not by the Lorentz field F, but by the cavity field G, carries out the explicit calculations for isotropic spherical molecules only, although he refers to the possible extension to ellipsoidal shapes. Allowance for anisotropy, i.e. for different polarizabilities in different directions within the still spherical molecule, has subsequently been made by Wilson (1939).

We may, on the other hand, take into account the departure from spherical shape of the solute molecules. We treat these as uniform ellipsoids, in which the charge distribution in the case of a vanishing internal field leads to a dipole moment μ in the direction of one of the axes. The ellipsoids are assumed to have an isotropic 'effective' refractive index n, defined, following Wilson, by

$$(n^2 - 1)M/(n^2 + 2)\rho = 4\pi N_{\mathrm{A}}\alpha/3 = {}_{\mathrm{D}}P, \qquad \ldots\ldots(9)$$

where ${}_{\mathrm{D}}P$ is the distortion polarization of the molecule (the sum of the electronic and atomic polarizations), α its mean polarizability, ρ the density of the pure substance at the temperature considered, M its molecular weight and N_{A} Avogadro's number. Ideally, ${}_{\mathrm{D}}P$ should be the value obtained from gas measurements; if such measurements are not available, one of the usual approximations such as ${}_{\mathrm{D}}P = 1{\cdot}05$ times $[M_{\mathrm{R}}]_{\mathrm{D}}$, the molecular refraction of the D-line, may be used.

For the treatment of dilute solutions in non-polar media, each ellipsoid is imagined surrounded by an isotropic continuum having the dielectric constant ϵ of the solvent (the limitations of this simplification have been discussed by Onsager (1936)). The external moment μ^* determining the force exerted at large distances on the surrounding liquid is then independent of higher moments in the multipole development of the charge distribution. If G is the cavity field parallel to the dipole under an applied field E in the same direction, Onsager's relation

$$\mu G = \mu^* E \qquad \ldots\ldots(10)$$

must hold irrespective of shape; for it does not matter whether we calculate the interaction energy as the energy of the dipole under the influence of the uniform field, or as the energy of the charges (imagined at large distances) producing that field in the potential of the dipole. This rule still applies if we do not remove the whole molecule inside its boundary, but, following the procedure of Fröhlich and Sack (1944), merely remove the dipole, leaving the polarizability unaltered. (Equation (10) requires modification if the media are no longer isotropic, for in that case the total interaction energy between the field and the induced polarization within two similar concentric surfaces no longer vanishes.) For a spherical cavity, G is given by

$$G=3\epsilon E/(2\epsilon+n^2),$$

and in the general case, according to (39), ($\epsilon_0=n^2$, $\epsilon_1=\epsilon$)

$$G=\frac{\epsilon}{\epsilon+(n^2-\epsilon)\xi}E \qquad \ldots\ldots(11)$$

which, with (10), leads to

$$\mu^*=\mu\epsilon/[\epsilon+(n^2-\epsilon)\xi]. \qquad \ldots\ldots(12)$$

Gas measurements correspond to $\epsilon=1$, in which case μ^* is the measured gas moment, μ_{gas};

$$\mu_{\text{gas}}=\mu/[1+(n^2-1)\xi], \qquad \ldots\ldots(13)$$

whence

$$\frac{\mu^*}{\mu_{\text{gas}}}=\frac{1+(n^2-1)\xi}{\epsilon+(n^2-\epsilon)\xi}\epsilon. \qquad \ldots\ldots(14)$$

A treatment exactly analogous to Onsager's leads to the same formula as the equation (42) of his paper (1936):

$$\epsilon_{\text{soln}}=n^2{}_{\text{soln}}+4\pi N_s(\mu^*)^2/3\,kT, \qquad \ldots\ldots(15)$$

where N_s denotes the number of solvent molecules per unit volume. The effective refractive index of the solution, n_{soln}, is obtained by adding the components of all the induced dipoles in the direction of the field; if λ_k denotes the angle between the field and the kth axis of the ellipsoid, and $\mu_E{}'$ is the component of the induced moment in the field, then (62), together with (39), (50) and (59), leads to

$$\begin{aligned}(n^2{}_{\text{soln}}-\epsilon)E/4\pi&=N_s[\mu_E{}']_{\text{av}}\\&=\tfrac{1}{3}N_s(n^2-\epsilon)\,abc\sum_{k=1}^{3}\left[\frac{\cos\lambda_k\,.\,E}{\epsilon+(n^2-\epsilon)\xi}\cos\lambda_k\right]_{\text{av}}.\end{aligned} \qquad \ldots\ldots(16)$$

On averaging, $\cos^2\lambda_k$ has to be replaced by $\frac{1}{3}$, and hence

$$n^2{}_{\text{soln}}-\epsilon=\tfrac{1}{3}v(n^2-\epsilon)\sum_{k=1}^{3}[\epsilon+(n^2-\epsilon)\xi_k]^{-1}, \qquad \ldots\ldots.(17)$$

where v denotes the volume fraction occupied by the solute. Strictly speaking, equation (9) should be replaced by

$$\frac{n^2-1}{9}\,\frac{M}{\rho}\sum_k\frac{1}{1+(n^2-1)\xi_k}={}_{\text{D}}P \qquad \ldots\ldots(18)$$

to make the model self-consistent; but the crude nature of some of the approximations involved, in particular the assumption of a uniform homogeneous n^2, hardly justifies the application of this additional correction.

If the apparent dipole moment μ_{soln} is determined by applying the Debye equation (4) to the value of ϵ_{soln} calculated with the use of (14), (15) and (17), the generalization of Onsager's formula

$$S_{\text{O}}=\mu_{\text{soln}}/\mu_{\text{gas}}=3\epsilon(n^2+2)/(n^2+2\epsilon)(\epsilon+2) \qquad \ldots\ldots(19)$$

is found to be

$$S_{RS} = \mu_{soln}/\mu_{gas} = \frac{3\epsilon[1-(n^2-1)\xi]}{(\epsilon+2)[\epsilon+(n^2-\epsilon)\xi]} \qquad \ldots\ldots(20)$$

or

$$\Delta\mu/\mu_{gas} = \frac{(\epsilon-1)[(\epsilon+2n^2)\xi-\epsilon]}{(\epsilon+2)[\epsilon+(n^2-\epsilon)\xi]}. \qquad \ldots\ldots(21)$$

For the special case when $\epsilon = n^2$, equation (20) reduces to the uncorrected Higasi formula (7).

§ 5. DISCUSSION AND COMPARISON WITH EXPERIMENTAL RESULTS

If the theory developed above is to be applied to actual substances, the dimensions of the ellipsoidal model representing the molecule must be known. A number of methods by which these dimensions may be estimated have been listed by Jenkins and Bauer (1936). These authors recommend the use of light scattering measurements on the pure (liquid) substance, but in view of the contention of Stuart and Volkmann (1933 a) and Mueller (1936) that the classical theory of light scattering cannot be applied to the liquid state, it is preferable to base the choice of axial ratios on the actual dimensions of the molecule. With flexible molecules, however, a variety of molecular shapes is possible, and polarizability measurements must be used, since by the use of such data the most probable molecular configurations can be decided upon. Because the three principal polarizabilities of a molecule are roughly proportional to its linear dimensions in the corresponding directions, the use of these to determine the axial ratios of the ellipsoid gives results not greatly different from those obtained from the geometrical shape of the molecule. To illustrate this, the theory of the solvent effect presented here is tested, using axial ratios calculated by both of these methods, although it should be emphasized that the consideration of the principal polarizabilities is not consistent with the use of an isotropic n^2 in the calculations of § 4. The axial ratios for a number of such molecules are plotted in Figure 2, where the values for

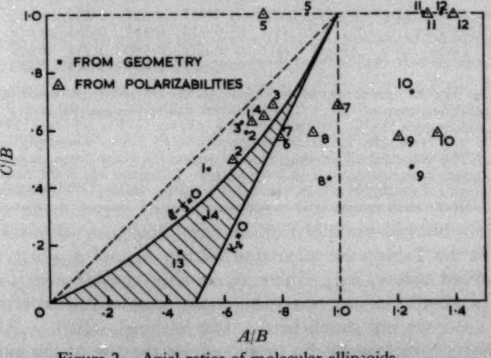


Figure 2. Axial ratios of molecular ellipsoids.
1. Ethyl ether ($10^{15} \times K_2$ at 760 mm. and 5400 A. $= -7{\cdot}6$). 2. sulphur dioxide ($-11{\cdot}1$). 3. methyl ether ($-5{\cdot}9$). 4. di-*iso*propyl ketone ($-18{\cdot}4$). 5. chloroform ($-8{\cdot}8$). 6. diethyl ketone ($-7{\cdot}6$). 7. acetone ($31{\cdot}1$). 8. pyridine ($21{\cdot}6$). 9. chlorobenzene ($29{\cdot}7$). 10. nitrobenzene (139). 11. methyl chloride ($35{\cdot}7$). 12. methyl bromide ($44{\cdot}3$). 13. 3 : 4 : 8 : 9-dibenzacridine (———). 14. 3 : 9-dimethylacridine (———).

the polarizability ellipsoid have been calculated from data by Stuart and Volkmann (1933 b), and the geometrical ratios have been evaluated using bond radii and angles given by Pauling (1944) and packing radii as given by Rees (1948).

Calculated and Observed Solvent Effects Solutions in Benzene ($\epsilon = 2{\cdot}226$) at 25° c.

						Molecular ellipsoid					
Substance	μ_{gas}	${}_{D}P$	ρ	n^2		A/B	C/B	ξ	S_{RS}	S_{H}	S_{obs}
Chlorobenzene	1·69	32·6*	1·101	2·405	G	1·251	0·463	0·181	*0·878*	*0·916*	*0·90*
					P	1·203	0·573	0·215	*0·908*	*0·935*	
Nitrobenzene	4·23	33·6*	1·489	3·055	G	1·255	0·727	0·236	*0·969*	*0·947*	*0·93*
					P	1·340	0·585	0·195	*0·927*	*0·924*	
Acetone	2·84	16·5*	0·784	1·860	G	0·805	0·600	0·315	*0·965*	*0·990*	*0·95*
					P	0·997	0·679	0·281	*0·937*	*0·971*	
Methyl ether	1·29^{a}	14·1*	0·654	1·751	G	0·666	0·628	0·368	*0·983*	*1·019*	—
					P	0·771	0·684	0·348	*0·967*	*1·008*	
Ethyl ether	1·18	33·2	0·708	1·855	G	(0·550	0·462	0·373	*0·998*	*1·022*)	*0·98*
					P	0·699	0·628	0·356	*0·984*	*1·013*	
Methyl chloride	1·86	15·4	0·906	2·145	G	1·298	1·000	0·267	*0·936*	*0·964*	*0·91*^{b}
					P	1·309	1·000	0·265	*0·934*	*0·963*	
Methyl bromide	1·78	19·1	1·674	2·522	G	1·356	1·000	0·257	*0·955*	*0·958*	—
					P	1·398	1·000	0·249	*0·948*	*0·954*	
Chloroform	1·02^{c}	22·5	1·480	2·160	G	0·892	1·000	0·364	*1·021*	*1·017*	*1·12*
					P	0·741	1·000	0·415	*1·065*	*1·046*	
Pyridine	—	25·3*	0·978	2·358	G	0·967	0·427	0·222	*0·912*	*0·939*	—
					P	0·912	0·487	0·253	*0·940*	*0·956*	
Sulphur dioxide	1·62^{d}	10·9	1·369	1·911	G	0·681	0·590	0·354	*0·988*	*1·012*	*1·00*^{d}
					P	0·634	0·497	0·349	*0·984*	*1·009*	

G and P denote that the dimensions of the ellipsoidal model have been calculated from geometry and from polarizabilities, respectively. The observed values of μ_{gas}, ${}_{D}P$ and S have been taken from (a) Groves and Sugden (1937), (b) Barclay and Le Fèvre (1950), (c) Goss (1940), (d) Le Fèvre and Ross (1950); the remainder are from the Appendix to *Trans. Faraday Soc.*, 1934, **30**, and Böttcher (1936). The values of ${}_{D}P$ are the observed gas values where these are reliable; the others, indicated by an asterisk, are calculated as $1{\cdot}05[M_R]_D$. Densities are from *International Critical Tables*, Vol. 3, except for methyl bromide, estimated from the value (1·7116) at 0°.

The results obtained on applying the new theory to various substances are summarized in the Table; the calculated values, S_{RS}, of μ_{soln}/μ_{gas} are compared with the observed values, S_{obs}, where reasonably reliable experimental results are available (it should be mentioned here that most of the experimental figures S_{obs} have an accuracy not much better than within $1-2\%$). Also tabulated, for comparison, are the results S_H predicted by the use of the amended Higasi equation (8).

It is seen that rather better agreement between theory and experiment is obtained from the new equation (20) than from (8), and that on the whole values of S calculated from polarizability data give better results than the geometrical values. The theory successfully predicts a small negative $\Delta\mu$ for sulphur dioxide and diethyl ether; methyl chloride, and especially chloroform, show solvent effects much greater than those calculable by any of the methods used.

On differentiating (20) with respect to ϵ, it is seen that the sign of $\partial S/\partial\epsilon$ is determined by the expression $(\epsilon^2+2n^2)\xi-\epsilon^2$. If $(1+2n^2)\xi<1$, then the apparent moment should always decrease with increase in ϵ; if $(1+2n^2)\xi>1$, the equation predicts an increase in μ_{soln} up to a maximum when $(\epsilon^2+2n^2)\xi=\epsilon^2$, followed by a decrease; for almost all substances this maximum will be reached at values of ϵ less than those of the normally available non-polar solvents, and hence μ_{soln} should decrease with increasing ϵ of solvent.

This has long been recognized with substances with negative $\Delta\mu$, but is contrary to the general belief when $\Delta\mu$ is positive; however, available data on the variation of the solvent effect with dielectric constant of the solvent are not accurate or numerous enough to support the general assumption; indeed, the results obtained for chloroform and bromoform by Müller (1933), who first drew attention to the solvent effect, if anything support our conclusion.

In view of the failure of Debye's theory to account correctly for the properties of polar liquids it may be pertinent to ask if the use of equation (4) should not be abandoned altogether. The equations (14) and (15) give a much more straightforward correlation between the molecular dipole moments and the macroscopic dielectric constant. Although the theory in its present state only applies to dilute solutions in non-polar media, with the additional assumption that the volumes are additive and that no strong short-range interaction is present between the molecules, an extension based on Onsager's approach appears to be promising enough to render unnecessary a re-interpretation of apparent dipole moments figuring in Debye's formula (4), such as given by equation (20).

§ 6. THE KERR CONSTANT RULE

A final point concerns the rule first stated by Goss (1934), and since frequently reiterated (cf. a review by Smith 1948), that the sign of the solvent effect, which according to the older theories is determined by $\xi - \frac{1}{3}$, should be opposite to that of the Kerr constant of the substance concerned. In fact this is not generally so, and the magnitude of the Kerr constant need bear no simple relation to the magnitude of the solvent effect.

The Kerr constant K of a molecule, a measure of its electric birefringence, is the sum of two terms K_1 and K_2; K_1, which depends on the molecular anisotropy, is positive and generally small; K_2, which involves the dipole moment as well, is generally very much larger than K_1, except for molecules with small or zero moments. The three principal polarizabilities of the molecule, α, β, γ, define its polarizability ellipsoid; if μ is parallel to one of the axes of this ellipsoid, say α, then K_2 is proportional to $\mu^2(2\alpha - \beta - \gamma)$; if α is the greatest or least axis of the ellipsoid, then K_2 will be respectively positive or negative; if α is the mid-axis, then K_2 may have either sign, and will be zero if $\alpha = \frac{1}{2}(\beta + \gamma)$.

If we assume that α, β, γ are proportional to the axes A, B, C of the geometrical ellipsoid (and it is on this not very accurate assumption that the rule is based), then $K_2 \sim \mu^2(2A - B - C)$, and we may attempt to correlate K_2 and $\xi - \frac{1}{3}$. This is done in Figure 2, in which the relation $K_2 = 0$ is represented by the heavy straight line, and $\xi - \frac{1}{3} = 0$ by the heavy curved line. Inspection of the figure shows that, for molecules lying in the region between these lines, shown shaded in Figure 2, K_2 and $\xi - \frac{1}{3}$ have the same (negative) sign. The contribution of the anisotropy term K_1 is usually small, and cannot be evaluated with reference to the polarizabilities alone; being positive, its effect will be that certain molecules lying close to, but to the left of, the line $K_2 = 0$, may yet have positive overall Kerr constants $K = K_1 + K_2$, but the general conclusion that there can arise cases in which K and $(\xi - \frac{1}{3})$ have the same sign, still stands, contrary to Goss' rule. In Figure 2 there are included also two molecules whose geometrical dimensions suggest that they might fail to comply with this Kerr constant rule. Further, it will be observed that, especially for molecules in which μ is parallel to the mid-axis, a large value of K does not necessarily imply a large value of $\xi - \frac{1}{3}$, and hence of the solvent effect.

With the use of the new theory the sign of the solvent effect according to (21) is determined by the expression $(\epsilon + 2n^2)\xi - \epsilon$; ξ alone is thus not sufficient to determine the sign of $\Delta\mu$, and the connection between $\Delta\mu$ and the Kerr constant becomes even more remote. The Table and Figure 2 show that sulphur dioxide and methyl chloride both show effects which are contrary to Goss' rule.

ACKNOWLEDGMENT

The authors wish to thank Professor R. J. W. Le Fèvre for helpful discussions on this topic and for permission to use unpublished results.

APPENDIX

CALCULATION OF THE FUNCTIONS ξ

In §2 it is pointed out that the general equations (1) for the internal field functions ξ are inconvenient for deriving numerical values for ξ in the case of ellipsoids with three unequal axes. A more convenient approach is found in the theory of ellipsoidal harmonics (Lamé functions) in association with Jacobian elliptic functions (cf., for example, Whittaker and Watson 1927).

If $a>b>c$, the appropriate modulus k, complementary modulus k', and base length l are

$$k=\left(\frac{a^2-b^2}{a^2-c^2}\right)^{1/2}; \quad k'=(1-k^2)^{1/2}=\left(\frac{b^2-c^2}{a^2-c^2}\right)^{1/2}; \quad l=(a^2-c^2)^{1/2} \qquad \ldots\ldots(22)$$

The Jacobian elliptic functions

$$\operatorname{sn} u=\operatorname{sn}(u,k), \quad \operatorname{cn} u=\operatorname{cn}(u,k), \quad \operatorname{dn} u=\operatorname{dn}(u,k) \qquad \ldots\ldots(23)$$

are introduced, defined by

$$\left(\frac{d\operatorname{sn} u}{du}\right)^2=(1-\operatorname{sn}^2 u)\ (1-k^2\operatorname{sn}^2 u); \quad \operatorname{sn} 0=0; \quad \left(\frac{d\operatorname{sn} u}{du}\right)_{u=0}=1 \qquad \ldots\ldots(24)$$

and

$$\operatorname{cn}^2 u=1-\operatorname{sn}^2 u, \quad \operatorname{dn}^2 u=1-k^2\operatorname{sn}^2 u; \quad \operatorname{cn} 0=\operatorname{dn} 0=1. \qquad \ldots\ldots(25)$$

The complete elliptic integrals of the first kind

$$\left.\begin{aligned} K=K(k)&=\int_0^{\pi/2}(1-k^2\sin^2\phi)^{-1/2}\,d\phi,\\ iK'(k)&=i\int_0^{\pi/2}(1-k'^2\sin^2\phi)^{-1/2}\,d\phi, \end{aligned}\right\} \qquad \ldots\ldots(26)$$

are the semi- or quarter-periods of the doubly periodic functions (23); the values of the functions for special values of u are

$$\operatorname{sn} K=1, \quad \operatorname{cn} K=0, \quad \operatorname{dn} K=k' \qquad \ldots\ldots(27)$$

$$\operatorname{sn}(K+iK')=k^{-1}; \quad \operatorname{cn}(K+iK')=-ik'/k; \quad \operatorname{dn}(K+iK')=0; \qquad \ldots\ldots(28)$$

whereas at $u=iK'$ all three functions have simple poles with residues k^{-1}, $-ik^{-1}$, $-i$ respectively.

The Cartesian coordinates x, y, z can now be expressed in the form

$$\left.\begin{aligned} x&=k^2 l\operatorname{sn}\alpha\operatorname{sn}\beta\operatorname{sn}\gamma,\\ y&=-(k^2/k')l\operatorname{cn}\alpha\operatorname{cn}\beta\operatorname{cn}\gamma,\\ z&=(i/k')l\operatorname{dn}\alpha\operatorname{dn}\beta\operatorname{dn}\gamma. \end{aligned}\right\} \qquad \ldots\ldots(29)$$

The surfaces upon which any one of the variables α, β, γ are constant are confocal quadrics; they are ellipsoids with semi-axes

$$a = kl \operatorname{sn} \alpha, \qquad b = ikl \operatorname{cn} \alpha, \qquad c = il \operatorname{dn} \alpha \qquad \ldots\ldots(30)$$

for constant α lying between iK' and $K+iK'$, hyperboloids of one sheet for β lying between K and $K+2iK'$, and hyperboloids of two sheets for γ lying between 0 and $4K$.

A harmonic function V, satisfying

$$\nabla^2 V = 0 \qquad \ldots\ldots(31)$$

can be separated into Lamé functions $V = \Lambda(\alpha) M(\beta) N(\gamma)$ where each of the functions satisfies an equation of the type

$$d^2\Lambda/d\alpha^2 = [n(n+1)k^2 \operatorname{sn}^2 \alpha + A]\Lambda, \qquad \ldots\ldots(32)$$

where n is the degree of the harmonic, and A a separation constant. In the case of a homogeneous field, $n=1$, and Λ is a multiple of $\operatorname{sn}\alpha$, $\operatorname{cn}\alpha$, or $\operatorname{dn}\alpha$, depending on the direction of the field. The introduction into a medium of dielectric constant ϵ_1 of an ellipsoid of dielectric constant ϵ_0 whose surface is given by the relation $\alpha = \alpha_0$ will introduce a disturbance depending mathematically upon the second solution of (32), viz.

$$\Lambda(\alpha) \int_{iK'}^{\alpha} \frac{du}{[\Lambda(u)]^2},$$

the lower limit being chosen as the pole of Λ as this corresponds to a falling off of the disturbance at large distances ($\alpha \to iK'$). Taking the case of the field acting in the direction of the *longest* axis, the potentials outside and inside the ellipsoid, V_e and V_i, can be written

$$\left.\begin{aligned} V_e &= -k^2 l\left(E - k^{-2} H_a \int_{iK'}^{\alpha} du/\operatorname{sn}^2 u\right) \operatorname{sn}\alpha \operatorname{sn}\beta \operatorname{sn}\gamma, \\ V_i &= -k^2 l G_a \operatorname{sn}\alpha \operatorname{sn}\beta \operatorname{sn}\gamma, \end{aligned}\right\} \qquad \ldots\ldots(33)$$

where G_a represents the cavity field in the ellipsoid, and $\frac{1}{3}H_a$ is a measure of the effective induced dipole. The quantities E, G_a, and H_a are related by the continuity conditions

$$V \text{ continuous at } \alpha = \alpha_0, \qquad \ldots\ldots(34)$$

$$(\epsilon \operatorname{grad} V)_{\text{normal}} \text{ continuous at } \alpha = \alpha_0, \qquad \ldots\ldots(35)$$

the latter relation being equivalent to

$$\epsilon \partial V/\partial\alpha \text{ continuous at } \alpha = \alpha_0. \qquad \ldots\ldots(36)$$

Substituting (34) and (36) into (33), we obtain

$$E - k^{-2} H_a \int_{iK'}^{\alpha} du/\operatorname{sn}^2 u - G_a = 0 \qquad \ldots\ldots(37)$$

$$\left[\epsilon_1\left(E - k^{-2} H_a \int_{iK'}^{\alpha_0} du/\operatorname{sn}^2 u\right) - \epsilon_0 G_a\right] d \operatorname{sn}\alpha_0/d\alpha_0 - \epsilon_1 H_a/k^2 \operatorname{sn}\alpha_0 = 0, \qquad \ldots\ldots(38)$$

and hence

$$G_a = \epsilon_1 E/[\epsilon_1 - (\epsilon_1 - \epsilon_0)\xi_a], \qquad \ldots\ldots(39)$$

$$H_a = G_a k^2 (\epsilon_1 - \epsilon_0) \operatorname{sn}\alpha_0 (d \operatorname{sn}\alpha_0/d\alpha_0)/\epsilon_1, \qquad \ldots\ldots(40)$$

where
$$\xi_a = -\operatorname{sn}\alpha_0\,(d\operatorname{sn}\alpha_0/d\alpha_0)\int_{iK'}^{\alpha_0} du/\operatorname{sn}^2 u. \qquad \ldots\ldots(41)$$

The difficulties arising from the complex values of α_0 can be eliminated by the transformation
$$\alpha' = \alpha_0 - iK' \qquad \ldots\ldots(42)$$
with (cf. Whittaker and Watson 1927)
$$\operatorname{sn}\alpha_0 = 1/(k\operatorname{sn}\alpha'). \qquad \ldots\ldots(43)$$
The equations (40) and (41) can then be expressed as
$$\begin{aligned}\xi_a &= -\frac{1}{\operatorname{sn}\alpha'}\frac{d}{d\alpha'}\left(\frac{1}{\operatorname{sn}\alpha'}\right)\int_0^{\alpha'}\operatorname{sn}^2 u\,du\\ &= \frac{1}{k^2}\frac{\operatorname{cn}\alpha'\operatorname{dn}\alpha'}{\operatorname{sn}^3\alpha'}\int_0^{\alpha'}(1-\operatorname{dn}^2 u)\,du\\ &= \frac{1}{k^2}\frac{\operatorname{cn}\alpha'\operatorname{dn}\alpha'}{\operatorname{sn}^3\alpha'}\left(\alpha' - \int_0^{\alpha'}\operatorname{dn}^2 u\,du\right), \end{aligned} \qquad \ldots\ldots(44)$$
and
$$H_a = \frac{\epsilon_0-\epsilon_1}{\epsilon_1}\frac{\operatorname{cn}\alpha'\operatorname{dn}\alpha'}{\operatorname{sn}^3\alpha'}G_a. \qquad \ldots\ldots(45)$$

Introducing the angle $\theta = \sin^{-1}(\operatorname{sn}\alpha')$, (44) and (45) become
$$\xi_a = \frac{1}{k^2}\frac{\cos\theta(1-k^2\sin^2\theta)^{1/2}}{\sin^3\theta}[F(k,\theta)-E(k,\theta)], \qquad \ldots\ldots(46)$$
$$H_a = G_a\frac{\cos\theta(1-k^2\sin^2\theta)^{1/2}}{\sin^3\theta}\,\frac{\epsilon_0-\epsilon_1}{\epsilon_1}, \qquad \ldots\ldots(47)$$
where $F(k,\theta)$ and $E(k,\theta)$ represent the elliptic integrals of the first and second kinds, respectively, and
$$\cos\theta = \operatorname{cn}\alpha' = i\operatorname{dn}\alpha_0/k\operatorname{sn}\alpha_0 = c/a. \qquad \ldots\ldots(48)$$

If the applied field is in the direction of the mid-axis, $2b$, of the ellipsoid, then the following equations replace equations (33) for the external and internal potentials:
$$\left.\begin{aligned} V_{\mathrm{e}} &= (k^2/k')l\left(E + k^{-2}H_b\int_{ik'}^{\alpha_0} du/\operatorname{cn}^2 u\right)\operatorname{cn}\alpha\operatorname{cn}\beta\operatorname{cn}\gamma,\\ V_1 &= (k^2/k')lG_b\operatorname{cn}\alpha\operatorname{cn}\beta\operatorname{cn}\gamma. \end{aligned}\right\} \qquad \ldots\ldots(49)$$
From the continuity conditions (34) and (36) we obtain
$$G_b = \frac{\epsilon_1 E}{\epsilon_1-(\epsilon_1-\epsilon_0)\xi_b}\,;\quad H_b = \frac{\epsilon_0-\epsilon_1}{\epsilon_1}\operatorname{cn}\alpha_0\frac{d\operatorname{cn}\alpha_0}{d\alpha_0}k^2G_b, \qquad \ldots\ldots(50)$$
where
$$\xi_b = -\operatorname{cn}\alpha_0\frac{d\operatorname{cn}\alpha_0}{d\alpha_0}\int_{iK'}^{\alpha_0} du/\operatorname{cn}^2 u. \qquad \ldots\ldots(51)$$
On substituting
$$\alpha_0 = K + iK' - \alpha'' \qquad \ldots\ldots(52)$$
with
$$\operatorname{sn}\alpha_0 = \frac{\operatorname{dn}\alpha''}{k\operatorname{cn}\alpha''}\,;\quad \operatorname{cn}\alpha_0 = \frac{k'}{ik\operatorname{cn}\alpha''}\,;\quad \operatorname{dn}\alpha_0 = \frac{k'\operatorname{sn}\alpha''}{i\operatorname{cn}\alpha''} \qquad \ldots\ldots(53)$$

we obtain for (51)

$$\xi_b = \frac{1}{\mathrm{cn}\,\alpha''}\,\frac{d}{d\alpha''}\left(\frac{1}{\mathrm{cn}\,\alpha''}\right)\int_{\alpha''}^{K}\mathrm{cn}^2 u\,du = \frac{\mathrm{sn}\,\alpha''\,\mathrm{dn}\,\alpha''}{k^2\,\mathrm{cn}^3\alpha''}\int_{\alpha''}^{K}(\mathrm{dn}^2 u - k'^2)\,du$$
$$= \frac{\sin\phi(1-k^2\sin^2\phi)^{1/2}}{k^2\cos^3\phi}[E(k)-E(k,\phi)-k'^2K(k)+k'^2F(k,\phi)], \quad \ldots\ldots(54)$$

where
$$\sin\phi = \mathrm{sn}\,\alpha'' = \mathrm{dn}\,\alpha_0/k\,\mathrm{cn}\,\alpha_0 = c/b, \quad \ldots\ldots(55)$$

and $E(k)=E(k,\pi/2)$ represents the complete elliptic integral of the second kind. For (50) we obtain from (53) and (55)

$$H_b = \frac{\epsilon_0-\epsilon_1}{\epsilon_1}\,\frac{\mathrm{sn}\,\alpha''\,\mathrm{dn}\,\alpha''}{\mathrm{cn}^3\alpha''}k'^2G_b = \frac{\epsilon_0-\epsilon_1}{\epsilon_1}\,\frac{\sin\phi(1-k^2\sin^2\phi)^{1/2}}{\cos^3\phi}k'^2G_b. \quad \ldots\ldots(56)$$

Finally, if the external field is in the direction of the minor axis, $2c$, of the ellipsoid, the equations equivalent to (33) and (49) are

$$\left.\begin{aligned} V_e &= -(i/k')\left[E + H_c\int_{iK'}^{\alpha} du/\mathrm{dn}^2 u\right]\mathrm{dn}\,\alpha\,\mathrm{dn}\,\beta\,\mathrm{dn}\,\gamma, \\ V_i &= -(i/k')G_c\,\mathrm{dn}\,\alpha\,\mathrm{dn}\,\beta\,\mathrm{dn}\,\gamma, \end{aligned}\right\} \quad \ldots\ldots(57)$$

which, with the boundary conditions (34) and (36), and the substitutions (52), (53) and (55), leads to

$$\xi_c = \frac{\mathrm{sn}\,\alpha''}{\mathrm{cn}\,\alpha''}\,\frac{d}{d\alpha''}\left(\frac{\mathrm{sn}\,\alpha''}{\mathrm{cn}\,\alpha''}\right)\int_{\alpha''}^{K}\frac{\mathrm{cn}^2 u}{\mathrm{sn}^2 u}du$$
$$= \frac{\mathrm{sn}\,\alpha''\,\mathrm{dn}\,\alpha''}{\mathrm{cn}^3\alpha''}\int_{\alpha''}^{K}\left[-\mathrm{dn}^2 u - \frac{d}{du}\left(\frac{\mathrm{cn}\,u\,\mathrm{dn}\,u}{\mathrm{sn}\,u}\right)\right]du = \frac{\mathrm{dn}^2\alpha''}{\mathrm{cn}^2\alpha''} - \frac{\mathrm{sn}\,\alpha''\,\mathrm{dn}\,\alpha''}{\mathrm{cn}^3\alpha''}$$
$$\times\int_{\alpha''}^{K}\mathrm{dn}^2 u\,du = \frac{1-k^2\sin^2\phi}{\cos^2\phi} - \frac{\sin\phi(1-k^2\sin\phi)^{1/2}}{\cos^3\phi}[E(k)-E(k,\phi)].$$
$$\ldots\ldots(58)$$

Also

$$G_c = \frac{\epsilon_1 E}{\epsilon_1-(\epsilon_1-\epsilon_0)\xi_c}\ ;\ H_c = k'^2\frac{\epsilon_0-\epsilon_1}{\epsilon_1}\,\frac{\sin\phi(1-k^2\sin^2\phi)^{1/2}}{\cos^3\phi}G_c. \quad \ldots\ldots(59)$$

Alternatively, one of the quantities ξ_a, ξ_b, ξ_c could have been derived from the other two by the use of (3).

Finally, to complete the treatment, we restate here the formulae for ξ for spheroidal molecules (Maxwell 1873).

(i) Prolate spheroid, $a>b=c$, $e=(1-b^2/a^2)^{1/2}$,

$$\xi_a = \left(\frac{1}{e^2}-1\right)\left(\frac{1}{2e}\ln\frac{1+e}{1-e}-1\right), \quad \xi_b=\xi_c=\tfrac{1}{2}(1-\xi_a). \quad \ldots\ldots(60)$$

(ii) Oblate spheroid, $a=b>c$, $e=(1-c^2/a^2)^{1/2}$,

$$\xi_c = 1/e^2-(1-e^2)^{1/2}\sin^{-1}e/e^3, \quad \xi_a=\xi_b=\tfrac{1}{2}(1-\xi_c). \quad \ldots\ldots(61)$$

The obstacle produces a field at large distances which, with the use of (29), (30), (33), (47), (49), (56), (57) and (59) is found equal to that of a dipole of strength

$$\mu' = \tfrac{1}{3}(\epsilon_0-\epsilon_1)\,abcG \quad \ldots\ldots(62)$$

in the direction of the applied field, in agreement with Maxwell's result.

The values of ξ_a, ξ_b, ξ_c as functions of the ratios $a : b : c$ have been calculated by first evaluating them for given values of k and ϕ (or θ), using four-figure tables of the circular functions and elliptic integrals, and interpolating for definite values of ξ (0·1, 0·2, . . . , 0·9). The interpolated values of k and ϕ (or θ) were then used to calculate the corresponding ratios $a : b : c$, and the results are shown in Figure 1. The accuracy obtained does not justify the presentation of the results in tabular form.

REFERENCES

Barclay, G., and Le Fèvre, R. J. W., 1950, *J. Chem. Soc.*, 556.
Böttcher, C. J. F., 1936, *Physica*, **6**, 59.
Debye, P., 1929, *Polar Molecules* (New York : The Chemical Catalog Co.), Chapter 3.
Frank, F. C., 1935, *Proc. Roy. Soc.* A, **152**, 171.
Fröhlich, H., and Sack, R., 1944, *Proc. Roy. Soc.*, A, **182**, 388.
Goss, F. R., 1934, *J. Chem. Soc.*, 696; 1937, *Ibid.*, 1915; 1940, *Ibid.*, 752.
Groves, L. G., and Sugden, S., 1937, *J. Chem. Soc.*, 1779.
Higasi, K., 1936, *Sci. Pap. Inst. Phys. Chem. Res., Tokyo*, **28**, 284.
Jenkins, H. O., and Bauer, S. H., 1936, *J. Amer. Chem. Soc.*, **58**, 2435.
Le Fèvre, C. G., and Le Fèvre, R. J. W., 1935, *J. Chem. Soc.*, 1747.
Le Fèvre, R. J. W., and Ross, I. G., 1950, *J. Chem. Soc.*, 283.
Maxwell, J. C., 1873, *Electricity and Magnetism* (Oxford : Clarendon Press), Section 437.
Mueller, Hans, 1936, *Phys. Rev.*, **50**, 547.
Müller, Horst, 1933, *Phys. Z.*, **34**, 689.
Onsager, L., 1936, *J. Amer. Chem. Soc.*, **58**, 1486.
Pauling, L., 1944, *Nature of the Chemical Bond* (Cornell : University Press).
Raman, C. V., and Krishnan, K. S., 1928, *Proc. Roy. Soc.*, A, **117**, 589.
Rees, A. L. G., 1948, *J. Chem. Phys.*, **16**, 995.
Smith, J. W., 1948, *Science Prog.*, **36**, 483.
Stuart, H. A., and Volkmann, H., 1933 a, *Z. Phys.*, **83**, 461; 1933 b. *Ann. Phys., Lpz.*, [v], **18**, 121.
Whittaker, E. T., and Watson, G. N., 1927, *Modern Analysis* (Cambridge : University Press), Chapters 22, 23.
Wilson, J. N., 1939, *Chemical Reviews*, **25**, 377.

A Study of the Horizontal Irregularities of the Ionosphere

By B. H. BRIGGS and G. J. PHILLIPS
Cavendish Laboratory, Cambridge

Communicated by J. A. Ratcliffe; MS. received 5*th April* 1950

ABSTRACT. The theory of diffraction by a random screen developed by Booker, Ratcliffe and Shinn is presented in a convenient form for practical application in ionospheric experiments. It is shown that measurements of the correlation of the fading of the reflected wave observed at spaced receiving points can be used to find the extent of the angular spreading of the downcoming wave.

Histograms are given to show the frequency of occurrence of different degrees of angular spreading observed during a series of experiments using pulse transmissions at vertical incidence.

For a frequency of 2·4 Mc/s. it is most common to find that the downcoming wave has an angular spread such that the amplitude falls to half value at an angle of 5° for regions E and F. For region F observed on 4·8 Mc/s., the corresponding value is 2·5°. There is no evidence for any pronounced seasonal or diurnal variations.

§ 1. INTRODUCTION

When a radio wave is returned from the ionosphere, the amplitude of the reflected wave is found to vary in a random way, and the wave is said to 'fade'. This fading is often dissimilar at two receiving points spaced about one wavelength apart on the ground (Ratcliffe and Pawsey 1933, Pawsey 1935). From this fact it may be deduced that the reflected wave must consist of a cone of rays spread over a considerable range of angles. In this paper we are concerned only with reflection at normal incidence. For this case, simple arguments used by Ratcliffe and Pawsey showed that the deviation of the extreme rays was of the order of 15° from the vertical for the normal E and F layers, and somewhat greater for the abnormal-E layer. The methods used, however, gave only the order of magnitude of the effects, because there was at the time no statistical theory which could be used to make accurate deductions from the observations. Such a theory has recently been developed by Booker, Ratcliffe and Shinn (1950) (we will refer to this paper as B.R.S.). These authors have considered the theoretical problem of diffraction by a random screen such as the ionosphere, using the auto-correlation functions to describe the properties of the random variables which arise.

The first part of the present paper is theoretical in nature, and aims at presenting the treatment of B.R.S. in a form which is convenient for practical applications. In § 2, we first summarize some of the results obtained by these authors. These results apply to the case of a plane wave incident normally on a random diffracting screen, which varies in one direction only—a 'one-dimensional' screen. In § 3, a standard form is assumed for the function which gives distribution of received energy as a function of the angle from the normal to the screen. In this way distributions of varying 'width' are represented by a single analytical expression with one variable parameter. The expected correlation between the amplitude at two points at a given distance apart is evaluated for this case, using the methods of B.R.S. An alternative method of calculation is then given, and is shown to give the same results. In § 4 the new method is used to extend the analysis to cover the case in which a spherical wave from a point source falls on a two-dimensional random screen. These results are directly applicable to the ionospheric experiments.

In § 5 it is shown that the angular spread of received energy can be deduced from the phase differences at spaced receiving points, instead of from the amplitude differences discussed earlier. Now the apparent direction of arrival of the reflected wave as indicated by a radio direction finder is directly related to these phase differences, and so the average departure of this direction from the mean direction may be related to the angular spread of energy in the reflected wave.

In § 6 we show how, subject to certain definitions and assumptions, it is possible to deduce a 'size' for the irregularities in the screen which are responsible for the diffraction process. The advantage of this method of expressing the results is that the size of the irregularities as defined in § 6 is theoretically independent of wavelength for a diffracting screen of given characteristics, unlike the angular spectrum which, for a given screen, has a width inversely proportional to the frequency.

In § 7 experimental results are given. These are obtained entirely from measurements of the correlation of signal amplitude. The experimental arrangement is one in which the fading of an echo returned from the ionosphere is observed at three receiving points placed at the corners of a right-angled triangle as described by Mitra (1949).

Observations on echoes from Regions E, F, and E_s have been made on many occasions (both day and night) over a period of one year, and statistical data are given for each region.

The variation of the angular spectrum with frequency has been verified by comparing the statistical results obtained on two frequencies, 2·4 Mc/s. and 4·8 Mc/s., and we have also shown that the size of the irregularities is, on the average, the same on the two frequencies.

§ 2. SUMMARY OF THEORETICAL TREATMENT OF A ONE-DIMENSIONAL RANDOM SCREEN

In this section we summarize some of the results obtained by B.R.S.

Consider a plane wave falling on a random screen which is one-dimensional, i.e. varying only in the X-direction.

The incident wave may be polarized with its electric vector either along or at right angles to the X-axis (Figures 1 (a) and (b)). The two cases may be treated in a symmetrical manner by working in terms of the magnetic field component H_y of the reflected wave in the first case, and the electric field component E_y in the second case.

These components may be expressed as the sum of contributions from an angular spectrum* of plane waves by means of the following equations in which $s \equiv \sin\theta$:

$$H_y = \eta \int_{-\infty}^{\infty} A_1(s) \exp\left(2\pi i \frac{sx - cz}{\lambda}\right) \frac{ds}{\lambda}, \qquad \ldots\ldots(1)$$

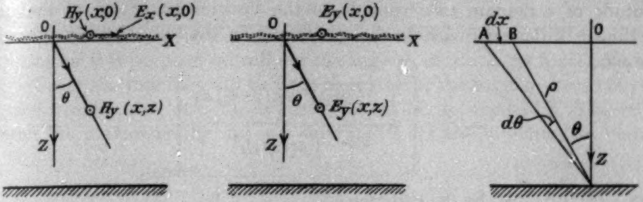
$$E_y = \int_{-\infty}^{\infty} A_2(s) \exp\left(2\pi i \frac{sx - cz}{\lambda}\right) \frac{ds}{\lambda}, \qquad \ldots\ldots(2)$$

---

* The spectral function employed here is slightly different from that in B.R.S., so that $A_1(s) \cos\theta$ would be equivalent to their $P(s)$.

where η is the ratio of magnetic to electric field intensities for a plane wave in free space.

To find the distribution of power in the angular spectrum consider the power received in a detecting device which responds to a small range of angles s to $s+\Delta s$.



(*a*) (*b*) (*c*)

Figure 1. Diagrams to show the notation used for the field components and the geometry of a random screen which varies only in the direction of the X-axis.

We have to perform the above integrations between the limits s and $s+\Delta s$. Assuming this range is large enough for the functions $A_1(s)$ and $A_2(s)$ to perform many random fluctuations the integration amounts to the summation of a number of complex quantities whose phases are random. Now we are not interested in the precise amplitude and phase of the field intensities for a given case. We wish, rather, to know their magnitude as a time average over a number of successive occasions for which the spectrum functions may have changed in detail but not in statistical properties. Taking such an average (which we shall denote by a bar) it can be shown that for the range s to $s+\Delta s$

$$\overline{|H_y|^2} \propto \overline{|A_1(s)|^2}\,\Delta s,$$

$$\overline{|E_y|^2} \propto \overline{|A_2(s)|^2}\,\Delta s.$$

The power flux in the direction $s=\sin\theta$ is therefore proportional to $\overline{|A_1(s)|^2}\,\Delta s$ and $\overline{|A_2(s)|^2}\,\Delta s$ respectively. Alternatively, if we use a detecting device which accepts a small range of angles θ to $\theta+d\theta$ we will receive powers $\overline{|A_1(s)|^2}\cos\theta\,d\theta$ and $\overline{|A_2(s)|^2}\cos\theta\,d\theta$ in the two cases.

Now consider a two-dimensional screen illuminated by a linearly polarized plane wave. It seems reasonable to require that for a random screen, the power should be scattered similarly in the two planes in and at right angles to the direction of polarization. This requires that $\overline{|A_1(s)|^2}=\overline{|A_2(s)|^2}$. We will assume that the screen satisfies this condition, and in future either function will be denoted by $\overline{|A(s)|^2}$.

If in equations (1) and (2) we put $z=0$, we get the variation of the field components of the reflected wave just after it leaves the screen. These fluctuations must be characteristic of the irregularities in the screen itself. Further, $A(s)$ is simply the Fourier transform of the appropriate field component. It follows that the finer the structure in the screen, the wider the 'spread' of the function $A(s)$. Irregularities in the screen of the order of one wavelength correspond to $s=1$, and give rise to energy travelling at 90° to the normal. Irregularities smaller than this give $|s|>1$ and give rise to evanescent waves which can be neglected at distances of more than several wavelengths from the screen.

In the ionospheric experiments it is possible to observe the *amplitude* of a field component in a horizontal plane (z = constant) as a function of x. A theorem due to Uhlenbeck relates the auto-correlation function of the fluctuation of amplitude of a random *time* variable to the Fourier transform of the power spectrum. With an appropriate change of variable, the same theorem gives from (1) or (2)

$$\rho_R(\xi) = \left[\frac{\int_{-1}^{+1} \overline{|A(s)|^2} \cos(2\pi s \xi)\, ds}{\int_{-1}^{+1} \overline{|A(s)|^2}\, ds} \right]^2 \qquad \text{......(3)*}$$

where $\rho_R(\xi)$ is defined by the equation

$$\rho_R(\xi) = \frac{\overline{R(x)R(x+\xi)} - [\overline{R(x)}]^2}{\overline{[R(x)]^2} - [\overline{R(x)}]^2} \qquad \text{......(4)}$$

R is the amplitude of the field component E_y or H_y and the bars denote averaging over x. The function $\rho_R(\xi)$ measures the average correlation between values of R measured at a distance $\lambda\xi$ apart (i.e. ξ is the distance measured in terms of the wavelength as unit).

By placing the limits in the integrals in (3) equal to ± 1 rather than $\pm\infty$ we have omitted the contribution of any evanescent waves in the angular spectrum. This result holds in general for observations made at a distance of many wavelengths from the screen. It holds at all distances if no evanescent waves are present, i.e. if the screen contains no structure comparable with the wavelength. Under these conditions, the function $\rho_R(\xi)$ describing the intensity pattern on any plane parallel to the screen is the same at all distances, and it will therefore be representative of the irregularities in the screen itself.

It is important to notice that the form of $\overline{|A(s)|^2}$ cannot be deduced uniquely from a knowledge of $\rho_R(\xi)$. This is because the upper integral in (3) may be negative for some values of ξ, and it is therefore not permissible to take the square root of each side of the equation and deduce that $\overline{|A(s)|^2}$ is the Fourier transform of $\sqrt{\rho_R(\xi)}$. This difficulty can be overcome by assuming a standard analytical form for the function $\overline{|A(s)|^2}$ which is such that a family of curves of varying angular spread are obtained by the variation of a single parameter. The corresponding family of curves for $\rho_R(\xi)$ can then be deduced from (3), and the member of this family which best fits the experimental results can be determined. This method is developed in §§ 3 and 4.

§ 3. A STANDARD FORM OF ANGULAR DISTRIBUTION

The function $\cos^n \theta$ will be used to represent a family of curves of varying angular spread (n being a variable parameter). This function is convenient because various correction factors of the same form appear in the calculations, as shown later. It may readily be shown that for n large (the case which will

---

[*] If the quantity inside the large brackets in (3) is denoted by ω, then a more accurate expression for $\rho_R(\xi)$ is $\pi\{\omega^2 + \omega^4/16\}/4(4-\pi)$ (Burgess and Uhlenbeck, unpublished, see B.R.S.). Equation (3) as given is accurate to within 10% over the range of ω from 0 to 1.

mainly be of interest) the function approaches a Gaussian distribution of standard deviation $1/\sqrt{n}$. Thus

$$\cos^n\theta \rightarrow \exp(-\tfrac{1}{2}n\theta^2) \qquad (n \text{ large}, -\tfrac{1}{2}\pi<\theta<+\tfrac{1}{2}\pi). \qquad \ldots\ldots(5)$$

It follows that $\cos^n 2\theta \simeq \cos^{4n}\theta$ for $-\tfrac{1}{4}\pi<\theta<+\tfrac{1}{4}\pi$; this result will be required later.

It is convenient to have a simple measure of the 'spread' of the function $\cos^n\theta$ and for this purpose we will use the angle θ_0 at which the function is equal to $\tfrac{1}{4}$. (Since the function will be used to represent the angular spread of *power*, this is the angle at which the *amplitude* has fallen to one-half.) The relation between the parameters θ_0 and n is thus $\cos^n\theta_0 = \tfrac{1}{4}$ and is plotted in Figure 2.

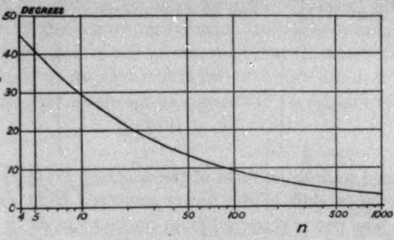


Figure 2. Relation between the two parameters θ_0 and n, as given by the equation $\cos^n\theta_0 = \tfrac{1}{4}$.

When n is large, and the distribution is approximately Gaussian, we have $\theta_0 = 1{\cdot}67(1/\sqrt{n})$ i.e. the half-width for amplitude is a constant factor times the standard deviation $1/\sqrt{n}$.

Let the energy received on a detector responding to H_y or E_y and sensitive to the range of angles θ to $\theta+d\theta$ be $W(\theta)\,d\theta$, where $W(\theta)=\cos^n\theta$. Then since $s=\sin\theta$ so that $ds=\cos\theta\,d\theta$ we can express this energy as $\overline{|A(s)|^2}\,ds=(1-s^2)^{(n-1)/2}\,ds$, and hence, by equation (3),

$$\rho_R(\xi) \propto \left[\int_{-1}^{+1}(1-s^2)^{(n-1)/2}\cos(2\pi s\xi)\,ds\right]^2. \qquad \ldots\ldots(6)$$

This gives the auto-correlation function for amplitude as a function of distance over the ground for the assumed form of angular distribution of received power.

The same result will now be obtained by an alternative method. The energy received by a detector sensitive to angles between θ and $\theta+d\theta$ may be regarded as coming from the region $\mathrm{AB}=dx$ of the screen in Figure 1(c). Now $d\theta=(\cos\theta\,dx)/\rho$ and the received power is therefore

$$\frac{W(\theta)\cos\theta\,dx}{\rho}. \qquad \ldots\ldots(7)$$

This is an expression for the power received from the element dx of the screen. As expected, the power scattered is proportional to the size of the element, and the factor $1/\rho$ may be interpreted as the spreading of the cylindrical wave from the element. The remaining factor depending on angle is the 'scattering polar diagram' which must be attributed to the element.

Thus we may regard the screen as built up of a number of scattering elements, each having a certain scattering polar diagram, and randomly placed in a horizontal plane. The advantage of this approach arises when the screen is two-dimensional and the illumination is non-uniform, but we will first show that it gives a correct result in the simple case of uniform illumination of a one-dimensional screen by a plane wave.

Let the diffracting screen move horizontally with velocity V in the direction of the X-axis. The diffraction pattern observed on any horizontal plane sweeps past a fixed observing point with the same velocity, and the fluctuations with distance are converted into fluctuations with time. Let the auto-correlation function for the fluctuations in time be $\rho_R(\tau)$. This function will be related to the function $\rho_R(\xi)$ by the transformation $\lambda\xi = V\tau$. Let the time fluctuations have a power spectrum $W(f)$. Then we have the result (analogous to equation (3)) that

$$\rho_R(\tau) = \left[\frac{\int_{-\infty}^{+\infty} W(f)\cos(2\pi f\tau)\,df}{\int_{-\infty}^{+\infty} W(f)\,df} \right]^2 . \qquad \ldots\ldots(8)$$

We now proceed to find the form of the function $W(f)$ by a consideration of the Doppler frequency shift of the power scattered from each element of the moving screen. The power scattered from the element dx in Figure 1 (c) suffers a frequency shift f given by

$$f = \pm \frac{V}{\lambda}\sin\theta = \pm ks \qquad \ldots\ldots(9)$$

where $k = V/\lambda$. The power received from this element for the particular form of $W(\theta)$ postulated above is, from (7),

$$\cos^n\theta\cos\theta\,dx/\rho = \cos^{n-1}\theta\,ds.$$

Hence, from (9) we have

$$W(f) \propto \left(1 - \frac{f^2}{k^2}\right)^{(n-1)/2} df \qquad (-k < f < +k)$$

and thus

$$\rho_R(\tau) \propto \left[\int_{-k}^{+k} \left(1 - \frac{f^2}{k^2}\right)^{(n-1)/2} \cos(2\pi f\tau)\,df \right]^2 .$$

To obtain $\rho_R(\xi)$, we put $\tau = \lambda\xi/V$. If, also, we put $f/k = u$ and note that $k = V/\lambda$, the equation gives

$$\rho_R(\xi) \propto \left[\int_{-1}^{+1} (1 - u^2)^{(n-1)/2} \cos(2\pi u\xi)\,du \right]^2 .$$

This result agrees with equation (6) obtained by the direct method.

§ 4. EXTENSION TO A TWO-DIMENSIONAL SCREEN WITH NON-UNIFORM ILLUMINATION

The method of calculation by Doppler shifts will now be extended to the following problem which is much nearer to the one actually encountered in the experiments. Let a random screen be illuminated from a transmitting aerial which has a polar diagram $\cos^t\theta$ (for power). Let the screen be such that if illuminated with a *plane* wave, it would scatter power so that the amount received per unit solid angle by a detector varies with θ in a manner given by the function $W(\theta)=\cos^q\theta$. Let the reflected wave be received on an aerial having a polar diagram $\cos^r\theta$ (for power). Then the power received from the element $dx\,dy$ in Figure 3 (a) may be written

$$dW=\underbrace{\left\{\frac{\cos^t\theta\cos\theta}{\rho^2}\right\}}_{\substack{\text{power flux}\\ \text{on the element.}}}\underbrace{\{\cos^{4q}\theta\cos\theta\,dx\,dy\}}_{\substack{\text{scattering power}\\ \text{of the element.}\\ \text{(cf. equation (7)).}}}\underbrace{\{\cos^r\theta\}}_{\substack{\text{receiving}\\ \text{aerial}\\ \text{sensitivity.}}}\underbrace{\left\{\frac{1}{\rho^2}\right\}}_{\substack{\text{spherical spreading}\\ \text{of the wave from}\\ \text{the element.}}}.$$

The term in this equation which represents the scattering power of the element requires further explanation. Since the elements are illuminated obliquely, equation (7) cannot be used directly. It is not possible to predict rigorously the form of the scattering function from a knowledge of the function for normal illumination, but it will be assumed that the main effect is to turn the polar diagram through an angle θ so that its maximum lies in the direction of the geometrically reflected ray, as shown in Figure 3 (b). The response will therefore be given by writing 2θ instead of θ in the term $\cos^q\theta$, or, better, by replacing q by $4q$, which gives a function which is a good approximation to the function $\cos^q 2\theta$ in the range $-\frac{1}{4}\pi<\theta<+\frac{1}{4}\pi$, and which falls smoothly to zero when θ lies outside this range. Thus we obtain the term for the scattering power of the element in the form given above.

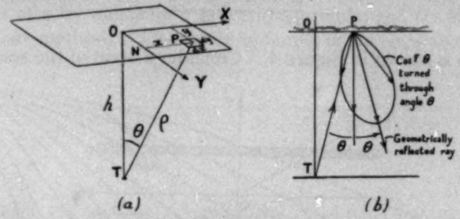


Figure 3.

(a) Diagram to show the notation used for a two-dimensional random screen.
(b) The polar diagram of an element of the screen when illuminated from a transmitter at T.

The power received from the element $dx\,dy$ may thus be written

$$dW=\cos^{(t+4q+r+6)}\theta\,dx\,dy \qquad \ldots\ldots(10)$$

since $1/\rho^4\propto\cos^4\theta$.

Now the solid angle subtended by the element is

$$d\Omega=\frac{\cos\theta\,dx\,dy}{\rho^2}\propto\cos^3\theta\,dx\,dy.$$

Thus the power may be written

$$dW\propto\cos^n\theta\,d\Omega, \qquad \ldots\ldots(11)$$

where

$$n=t+4q+r+3. \qquad \ldots\ldots(12)$$

Equation (11) gives the resultant angular distribution of received power. We now wish to relate this to the auto-correlation function of amplitude over the ground, as in the one-dimensional case. It may be noted that as far as the further development of the present section is concerned, it is possible to take equation (11) as the starting point, and thus avoid the approximations used above. Later, in § 6, however, we wish to deduce the value of q from the observed value of n, and the approximations cannot then be avoided.

The analysis will now be developed as in § 3. Let the diffracting screen move horizontally with velocity V in the direction of the X-axis. The diffraction pattern moves with velocity $2V$ for the mode of illumination here considered, so that the transformation from space to time auto-correlation functions is made by writing $\lambda\xi = 2V\tau$ in this case.

The energy returned from the element at P in Figure 3 (a) will have a Doppler frequency shift given by

$$f = \frac{2V}{\lambda} \cos \widehat{\mathrm{NPT}} = \frac{kx}{(x^2+y^2+h^2)^{1/2}},$$

where $k = 2V/\lambda$. Thus elements of the screen which produce a constant frequency shift f lie on the hyperbola

$$\frac{x^2}{f^2} = \frac{x^2+y^2+h^2}{k^2}. \qquad \ldots\ldots(13)$$

This hyperbola is shown in Figure 4. Consider a strip of the screen of width dy and distant y from the X-axis. From (13) we have

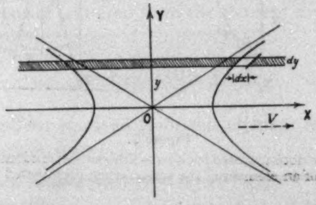


Figure 4. Hyperbola of constant Doppler frequency shift.

$$dx = \frac{(x^2+y^2+h^2)^{3/2}}{k(y^2+h^2)}\, df.$$

This equation gives the range dx within the strip corresponding to the frequency interval df. The power received from the element $dx\,dy$ is from (10) and (12)

$$dW \propto \frac{dx\,dy}{(x^2+y^2+h^2)^{(n+3)/2}}.$$

The contribution to the power spectrum for the frequency range f to $f+df$ is therefore

$$dW(f)\,df \propto \frac{dy\,df}{(x^2+y^2+h^2)^{n/2}(y^2+h^2)},$$

where (x, y) lies on the hyperbola appropriate to the frequency f, i.e. (x, y) satisfies equation (13). Using this fact we have

$$dW(f) \propto \frac{dy}{k(y^2+h^2)^{(n+2)/2}}\left(1 - \frac{f^2}{k^2}\right)^{n/2}.$$

Thus

$$W(f) \propto \left(1-\frac{f^2}{k^2}\right)^{n/2} \int_{-\infty}^{+\infty} \frac{dy}{(y^2+h^2)^{(n+2)/2}}.$$

The integral in this equation is a numerical factor of no interest. Thus

$$W(f) \propto \left(1-\frac{f^2}{k^2}\right)^{n/2}. \qquad \ldots\ldots(14)$$

This expression holds for $-k<f<k$, and $W(f)$ is zero for f outside these limits. From (8)

$$\rho_R(\tau) \propto \left[\int_{-k}^{+k} \left(1-\frac{f^2}{k^2}\right)^{n/2} \cos(2\pi f\tau)\, df\right]^2$$

and, writing $f/k=u$, $\lambda\xi=2V\tau$ and $k=2V/\lambda$, we have

$$\rho_R(\xi) \propto \left[\int_{-1}^{+1} (1-u^2)^{n/2} \cos(2\pi u\xi)\, du\right]^2. \qquad \ldots\ldots(15)$$

This expression gives the required auto-correlation function. It has been evaluated for small values of n by means of a reduction formula, and for large values of n by an approximate method. The resulting family of curves is shown in Figure 5.

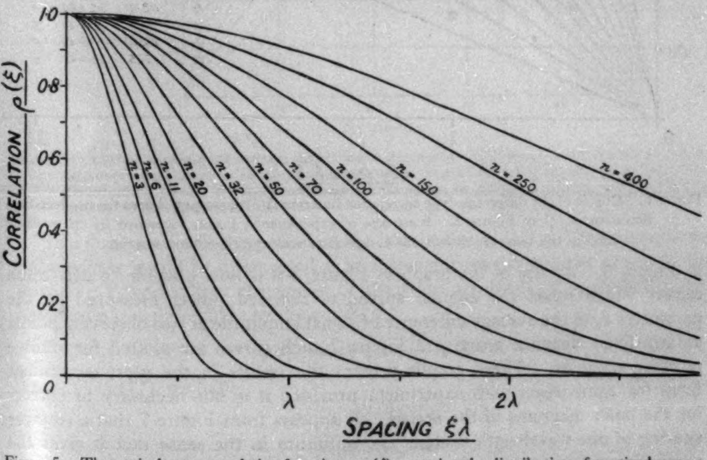


Figure 5. Theoretical auto-correlation functions $\rho_R(\xi)$ assuming the distribution of received power with angle to be given by the function $W(\theta)=\cos^n\theta$.

In practice the auto-correlation function $\rho_R(\xi)$ is inconvenient for computations, because it is based upon the average value of the product of $R(x)$ and $R(x+\xi)$ (equation (4)). It is more convenient to use a measure of correlation $\Delta(\xi)$ based on the average value of the *difference* of the amplitudes at the two points. An auto-correlation function of this type has been used in the study of random noise (Fürth and MacDonald 1947). We define

$$\Delta(\xi) = \overline{|R(x)-R(x+\xi)|}.$$

When it is necessary to refer to $\Delta(\xi)$ in words, we will call it the 'average difference' for brevity; strictly, it is the average value of the modulus of the difference.

It can be shown (McNicol 1949), subject to certain conditions which are usually satisfied in practice, that $\Delta(\xi)$ and $\rho_R(\xi)$ are related by the equation

$$\Delta(\xi) = 0{\cdot}59\overline{R}[1-\rho_R(\xi)]^{1/2}. \qquad \ldots\ldots(16)$$

By means of this equation, curves of $\Delta(\xi)/\overline{R}$ have been plotted, and are shown in Figure 6. By use of the graph in Figure 2 it is now possible to derive the curves which relate the angular spread of received power measured by the parameter θ_0 to the average difference of signal amplitude at two observing points at any fixed distance apart. In Figure 7 such curves are plotted for various receiver spacings. These graphs present the results in the most convenient form for comparison with experiment provided it is not necessary to correct for the polar diagrams of the aerials. It appears from Figure 7 that a receiver spacing of one wavelength is about the optimum in the sense that it gives the maximum sensitivity over the range of angles which is mainly of interest, and this spacing has been used most frequently in the experiments.

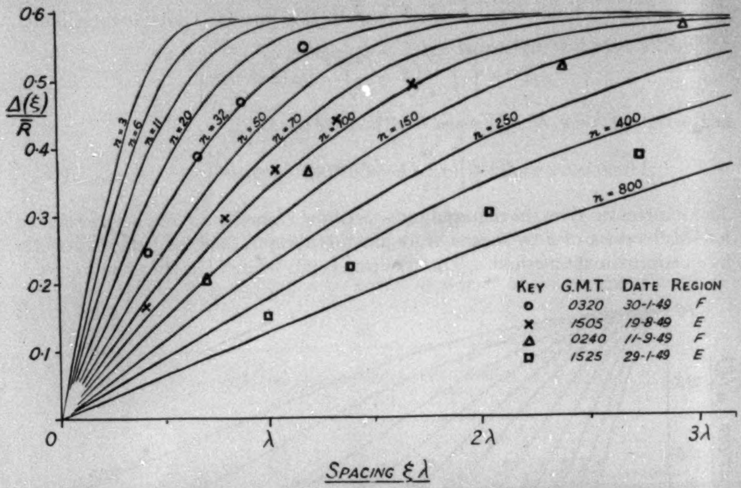


Figure 6. Curves of the difference-type correlation function $\Delta(\xi)$ corresponding to the product-type functions $\rho_R(\xi)$ of Figure 5. Four sets of experimental points, obtained by the method explained in the text, are shown, for comparison with the theoretical curves.

§ 5. PHASE DIFFERENCES AND RELATION TO DIRECTION FINDING EXPERIMENTS

The average difference of phase of the signals observed at a pair of spaced receivers will now be calculated. MacDonald (1949) has considered the analogous problem for a random time variable, and has shown that, if the fluctuation has a power spectrum of Gaussian form, with standard deviation σ, then

$$\overline{|\phi(t)-\phi(t+\tau)|} = 2\pi\sigma\tau \qquad \text{for } \tau \text{ small,}$$

where $\phi(t)$ is the instantaneous phase observed at time t. A comparison of equations (3) and (8) shows that the analogous equation for the fluctuation of phase over the ground is

$$\overline{|\phi(x)-\phi(x+\xi)|}=2\pi\sigma\xi$$

provided that $|A(s)|^2$ is Gaussian, with standard deviation σ. With the form for $|A(s)|^2$ assumed in § 3, this is the case for n large and $\sigma \simeq 1/\sqrt{n}=\theta_0/1{\cdot}67$. Thus

$$\overline{|\phi(x)-\phi(x+\xi)|}=2\pi\theta_0\xi/1{\cdot}67. \qquad \ldots\ldots(17)$$

This equation, then, gives the required average phase difference for a pair of receivers a distance ξ wavelengths apart, in terms of the parameter θ_0 specifying the spread of the angular spectrum.

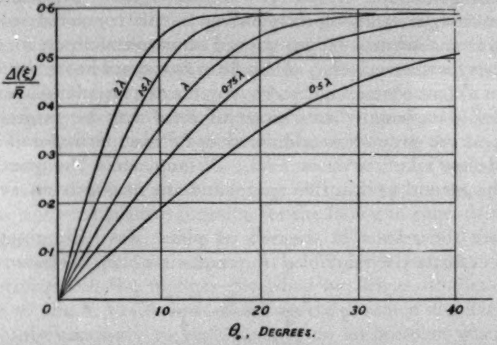


Figure 7. Curves showing the average difference of signal amplitude at a pair of receivers at various distances apart in terms of the parameter θ_0 specifying the degree of angular spreading of the downcoming wave. By means of these curves θ_0 can be obtained directly from the fading records at two such receivers.

In a direction-finding experiment, the instantaneous phase difference between the receivers would be interpreted as an instantaneous direction of arrival of the wave θ_i. Provided that the angles of deviation are small, θ_i is related to the observed phase differences by the equation

$$\phi(x)-\phi(x+\xi)=\frac{2\pi}{\lambda}\lambda\xi\theta_i.$$

Thus, from (17), we have

$$2\pi\xi\overline{|\theta_i|}=2\pi\theta_0\xi/1{\cdot}67$$

or

$$\overline{|\theta_i|}=0{\cdot}6\,\theta_0. \qquad \ldots\ldots(18)$$

This equation relates the *mean* bearing deviation (without regard to sign) to the half-width θ_0 of the angular spectrum. The extreme deviations observed, will, of course, be much greater than θ_0.

§ 6. CORRECTION FOR AERIAL POLAR DIAGRAMS AND DEDUCTION OF THE SIZE OF THE IRREGULARITIES

We have so far assumed the angular distribution of received power to be given by the function $\cos^n\theta$, and have considered the amplitude and phase differences at spaced receiving points which arise from this distribution. As a result it was possible to construct curves from which the parameter n or the closely related parameter θ_0 may be deduced from observations. That is to say n may be found by observing which curve in Figure 6 comes nearest to the point representing the observed correlation for a known receiver spacing, or which best fits a complete experimental correlation curve if this is available.

If however we want to go further than the determination of the actual received angular power density, and wish to express the amount of scattering by the ionosphere in a form independent of aerial properties, it is necessary to consider the separate contributions to the value of n as expressed by equation (12).

In the first place we can correct for the effect of aerial polar diagrams. This is only an important correction when the spreading of the downcoming wave is relatively large. The corrected value $n'=n-t-r$ would be obtained in an experiment using isotropic aerials for which $t=r=0$. The values of θ_0 given in the next section are those corresponding to this corrected value of n. The experimental arrangement usually employed a horizontal dipole for the transmitter and loop aerials for the receivers. A dipole in free space has a $\cos^2\theta$ polar diagram (for power) in a plane containing its own length and a circular diagram in the plane at right angles to its length; an average of $\cos\theta$ may be assumed. Owing to the presence of the ground an additional $\cos^2\theta$ is superimposed in both planes and we therefore take $\cos^3\theta$, or $t=3$. A loop aerial (magnetic dipole) will behave on the ground as if in free space, and we may take an average of $\cos\theta$, or $r=1$.

To correct the results to the case of plane wave illumination it is only necessary to evaluate the value of q from equation (12). Thus

$$q=\frac{n'-3}{4}=\frac{n-t-r-3}{4}$$

and the function $\cos^q\theta$ gives by the definition of q the angular distribution of received power which would be obtained if the ionosphere had plane waves incident upon it, and isotropic aerials were employed. This last step has taken into account the different range and obliquity factors involved in this case.

It is now possible to deduce something about the size of the ionospheric irregularities because we know that for plane wave illumination the autocorrelation function of the diffraction pattern on the ground is the same as that of the irregularities themselves, provided that no evanescent waves are present. We therefore define a characteristic length or 'size', l, of the irregularities as the distance between two points on the ground for which the correlation is 0·5, in the idealized case of plane wave illumination and 'all-round' aerials. (This is similar to a definition often used for the size of eddies in the theory of turbulence.) The value of l can be found from the curve in Figure 5 corresponding to the deduced value of q in place of n, by noting the spacing at which $\rho=0{\cdot}5$.

The main correction which involves a factor of four in the power of $\cos\theta$ is equivalent to assuming that the irregularities in the diffraction pattern actually produced are twice as large as those in the diffracting screen itself, because the illumination is from a point source at the ground rather than by a plane wave.

§7. EXPERIMENTAL RESULTS

The apparatus which has been used to study the fading at spaced receivers is in principle the same as that described by Mitra (1949). (Some improvements in experimental technique will be described elsewhere.) Three receivers are placed at the corners of a right-angled triangle, so that one pair lies in a N–S line and another pair in an E–W line, the spacing of each pair being approximately 130 metres. The fading of a single echo from a local pulse transmitter is recorded photographically at each point, the three records being recorded on a single film for ease of comparison.

It is first desirable to confirm that the assumed form of angular distribution provides a good approximation to those which occur in practice. This can be done by comparing the detailed shape of the theoretical auto-correlation functions $\Delta(\xi)$ of Figure 6 with experimentally determined curves. It would appear at first sight that in order to determine the latter, it would be necessary to have a large number of receivers at various spacings along a line, so that the correlation could be determined for a large number of spacings. Fortunately this is not so, for the following reason. If three receivers are placed at the corners of a right-angled triangle, it is possible by a comparison of the three fading records to detect any uniform movement of the diffraction pattern over the ground, and also to determine whether the pattern moves without change of form, or whether there are random changes present in addition to the steady movement (Briggs, Phillips and Shinn 1950). It is found that there are occasions when the random changes are negligible compared with those due to the steady drift. The fading observed at a fixed point is then due to an unchanging diffraction pattern moving over the ground, and the auto-correlation function for the fading in time, $\Delta(\tau)$, will be the same as the auto-correlation function for distance, $\Delta(\xi)$, subject to the transformation $\lambda\xi = V\tau$, where V is the drift velocity of the diffraction pattern. The velocity V can be determined in the manner described by Briggs, Phillips and Shinn. Thus, in order to obtain the desired check on the assumed angular distribution function, it is only necessary to evaluate $\Delta(\tau)$ on an occasion when the fading has been shown, from the three receiver experiment, to be due predominantly to a drifting diffraction pattern.

The method which has been used to compute $\Delta(\tau)$ in practice is as follows. A tracing of the fading curve is made and is placed over the original fading record with a time shift equal to the desired value of τ. The included area between the original and the shifted curve is measured with a planimeter. $\Delta(\tau)$ is equal to this area divided by the length of the record. The area underneath one fading record is also measured, and $\bar{R}$ is obtained by dividing this area by the length of the record. $\Delta(\tau)/\bar{R}$ is then plotted against τ. Some of the experimental points are shown in Figure 6 for comparison with the theoretical curves. It will be seen that the value of n can be determined for each example with some accuracy.

This method is restricted to occasions when the fading is due entirely to a movement of an unchanging diffraction pattern, and is also too laborious for general application. In general only a single point on the $\Delta(\xi)$ curve is known, obtained from the average difference between the fading records at the two receiving points. We have measured this quantity for a large number of records with a planimeter, and as a result of the work it has been found that an experienced observer is able to estimate the average difference between the records (without any measurement at all) to an accuracy of about 20%. For compiling statistical results this accuracy is sufficient, and in view of the very large number of records which have been obtained, it has been necessary to use this method for the largest proportion of the records, which were obtained from January 1949–January 1950 on a frequency of 2·4 Mc/s. Recently we have developed a method of recording the required average difference directly, and this method has been used to obtain further results on a frequency of 4·8 Mc/s. (January–March 1950).

From Figure 7 each record yields a value of θ_0, the angle to the half-amplitude point of the function representing the cone of received rays. In Figure 8 histograms* are given to show the number of occasions on which different values of θ_0 were observed for regions F, E, and E_s. A value of θ_0 can be obtained from the E–W pair of receivers and from the N–S receivers, and on the average the two values are found to be equal. though on individual occasions significantly different values may be obtained. In constructing the histograms the two values have been averaged. It will be seen that for 2·4 Mc/s., regions E and F give similar histograms with a value of θ_0 of about 5° most frequently observed. The histogram for region E_s has a less pronounced peak and also extends to larger values of θ_0. On this frequency the results on region E are entirely day-time observations, and the results on region F and region E_s are entirely night-time. On the frequency of 4·8 Mc/s., region F has been observed by both day and night. As predicted by the theory, the angular spread is smaller, and the peak in the histogram occurs at a value of θ_0 of about 2·5°. The number of occasions on which echoes from region E_s were observed on this frequency was too small to yield statistical results.

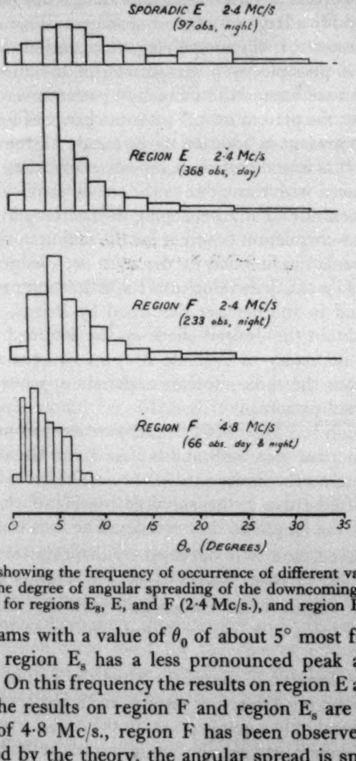


Figure 8. Histograms showing the frequency of occurrence of different values of the parameter θ_0, which specifies the degree of angular spreading of the downcoming wave. Separate histograms are shown for regions E_s, E, and F (2·4 Mc/s.), and region F (4·8 Mc/s.).

---

* Owing to the various ways in which the original data have been transformed and corrected, the histograms given in this paper have unequal intervals of the abscissae. In every case, the number of observations contained between two values of the abscissae in the final histogram has been divided by the length of the interval, so that the ordinates of the histograms are proportional to the number of observations per unit range of the abscissa.

The results shown in Figure 8 relate entirely to the properties of the cone of rays falling on the aerial system, and each histogram may be expected to hold only for the frequency on which it was obtained. If histograms are constructed for the sizes of the ionospheric irregularities deduced by the method described in §6, these should hold for all frequencies, except in so far as irregularities at different heights in the ionosphere with possibly different statistical properties may become effective as the frequency is changed.

The histograms of Figure 9 were obtained for a number of occasions on which irregularities of different sizes were observed. The results for the two frequencies 2·4 Mc/s. and 4·8 Mc/s. are seen to be in agreement.

When the fading records at the spaced receivers are highly correlated, it should be pointed out that the procedure used in deducing the size of the irregularities involves a considerable extrapolation of the data, and is dependent entirely on the particular shape postulated for the correlation functions. In essence, we are attempting to predict from the small measured difference between a pair of closely spaced receivers at what distance apart the receivers would have to be placed in order to obtain records almost uncorrelated. No high accuracy can therefore be claimed, though we have already given some indirect evidence which suggests that the assumed shape of the correlation functions is, in fact, obeyed up to quite large distances.

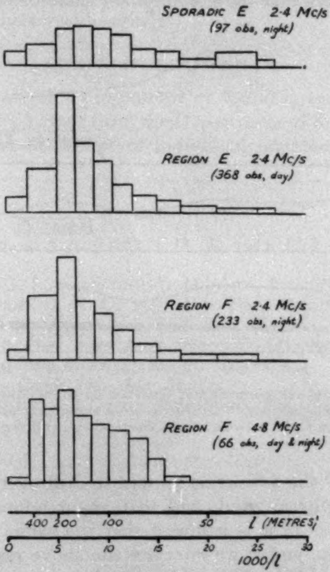


Figure 9. Histograms showing the frequency of occurrence of different values of the characteristic length l which specifies the size of the ionospheric irregularities. Separate histograms are shown for regions E_s, E, and F (2·4 Mc/s.) and region F (4·8 Mc/s.).

Some implications of the results may be noted. As the most frequent size of the irregularities is found to be of the order of 200 m. it follows that on radio wavelengths such that $\lambda > 200$ m. the ionosphere will behave as a 'completely rough' screen, as defined by B.R.S.

With the notation of § 4, such a screen would have $q = 0$, and if 'all-round' aerials were used, would have $n = 3$. The auto-correlation function would be given by the curve marked $n = 3$ in Figure 5. The fading on these wavelengths should, therefore, be practically uncorrelated at a receiver spacing of half a wavelength. For radio wavelengths such that $\lambda < 200$ m., however, the auto-correlation function becomes independent of wavelength, and therefore the correlation at one half wavelength spacing becomes greater as the wavelength decreases. In this range, receivers at a fixed absolute spacing tend to show the same correlation, independent of the wavelength. From the results of experiments with three receivers, interpreted as described by Briggs, Phillips and Shinn (1950), we believe that the predominant cause of fading in the short-wave range is a tendency for the diffraction pattern to drift over the ground due to the presence of an ionospheric wind, and that any random changes are only of secondary importance. If it is assumed that the same mechanism holds also in the range $\lambda > 200$ m., and if we interpret the above results in terms of speed of fading, we see that for $\lambda > 200$ m. we should expect the fading speed to be inversely proportional to wavelength, and for $\lambda < 200$ m. the fading speed should be independent of wavelength. The precise wavelength at which the change-over occurs, may, of course, be different on different occasions. These consequences are at present being tested by more direct experiments. As vertical incidence has been used in the experiments the conclusions of this paragraph will not in general apply to oblique incidence.

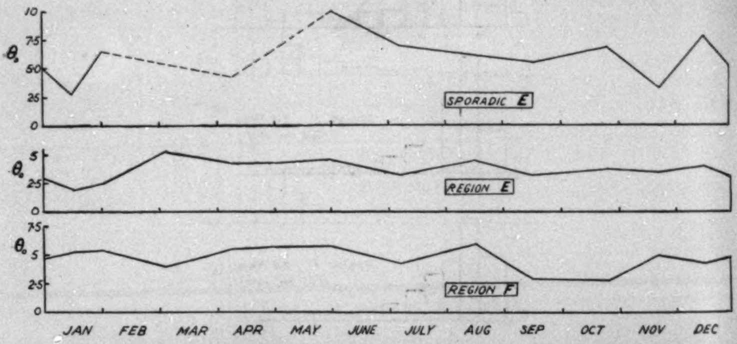


Figure 10. Average values of the parameter θ_0 observed on three days of each month through the year (1949–50), for a frequency of 2·4 Mc/s. No significant seasonal variation is apparent. The results for region E relate to day-time observations, and the results for regions F and E_s to night-time observations.

It is of interest to know whether the measurements show any important diurnal or seasonal variations. Results extending over a year are available for a frequency of 2·4 Mc/s. Figure 10 shows the average value of the parameter θ_0 obtained for each month of the year 1949–1950, for the three regions, E, F, and E_s.

No significant seasonal variation is apparent, nor has any significant diurnal variation been found. The diurnal variation on this frequency is complicated by the changing time at which the transition from region E to region F occurs at different times of the year. However, it can be stated that any diurnal variation is much smaller than the day to day and hour to hour variability.

Two limitations to the present theory should be noted: (i) When the scattering is very extreme, the received pulse is appreciably spread in range and no unique value can be assigned to its amplitude. This effect occurs mainly near the penetration frequencies of the regions, and also occasionally during the later part of the night when the 'spread F' phenomenon occurs. (ii) If the fading is 'shallow' (displaced Gaussian rather than Rayleigh amplitude distribution) there must be an appreciable specular component (McNicol 1949). This corresponds to a line in the angular spectrum at $\theta = 0$. The assumed form of angular spectrum then breaks down and the results given in the present paper are inapplicable. We have rejected a small proportion of records which come under one or other of these two headings.

ACKNOWLEDGMENTS

This work forms part of a programme of Radio Research at the Cavendish Laboratory supported by a grant from the Department of Scientific and Industrial Research. We are also in receipt of individual maintenance grants from the same Department.

REFERENCES

Booker, H. G., Ratcliffe, J. A., and Shinn, D. H., 1950, *Phil. Trans. Roy. Soc.* A, **262,** 579.

Briggs, B. H., Phillips, G. J., and Shinn, D. H., 1950, *Proc. Phys. Soc.* B, **63,** 106.

Fürth, R., and MacDonald, D. K. C., 1947, *Proc. Phys. Soc.*, **59,** 388.

MacDonald, D. K. C., 1949, *Proc. Camb. Phil. Soc.*, **45,** 368.

McNicol, R. W. E., 1949, *Proc. Instn. Elect. Engrs.*, Pt. III, **96,** 517.

Mitra, S. N., 1949, *Proc. Instn. Elect. Engrs.*, Pt. III, **96,** 441.

Pawsey, J. L., 1935, *Proc. Camb. Phil. Soc.*, **31,** 125.

Ratcliffe, J. A., and Pawsey, J. L., 1933, *Proc. Camb. Phil. Soc.*, **29,** 301.

Periodic Fading of Short-Wave Radio Signals

By S. R. KHASTGIR and P. M. DAS
Benares Hindu University, Benares, India

MS. received 23rd September 1949, *and in final form 26th April* 1950

ABSTRACT. Periodic fading patterns were recorded photographically with Calcutta signals of frequency 4,840 kc/s. received at Dacca (distance 240 km.) during the evening and early night hours of December 1948 and January 1949. The main features in the experimental conditions were : (i) the operating frequency was much less than the maximum usable frequency (M.U.F.) for the F layer transmission, (ii) the frequency was slightly greater than the M.U.F. for the ordinary wave transmission through the E layer and (iii) it was slightly less than the M.U.F. for the extraordinary wave transmission through the E layer between the transmitting and receiving stations. The following patterns of periodic or rhythmic fading were observed :

(i) Sinuous fading of comparatively quick period : this is considered to be of magneto-ionic origin, due to the interference between the upper and lower trajectory extraordinary waves in the E layer, the ordinary waves having passed through the E layer.

(ii) Periodic or rhythmic fading of comparatively slow period : the slow periodic fading is considered to be due to the beat-effect between the singly and doubly reflected waves from the F2 region or between the singly reflected waves from the E and F2 regions, the two interfering waves in different directions having suffered different amounts of Doppler change of frequency due to the vertical movement of the ionospheric layer or layers. The vertical velocity of the ionosphere as computed from this view agrees with the observed value.

(iii) Slow periodic fading with superposed ripples : this was observed when the ionospheric conditions were favourable for the simultaneous occurrence of the magneto-ionic type of sinuous fading and the Doppler beat type of slow periodic fading. In a few patterns of periodic fading there was evidence of extremely high frequencies (4–12 cycles/second) the origin of which is unknown.

§ 1. INTRODUCTION

THE work of Appleton and Beynon (1947) has shown that the interference of the ordinary and extraordinary components of a short-wave signal caused by magneto-ionic splitting in the ionosphere gives rise to rhythmic fading at a time, usually during the evening or early morning hours, when there is a steady decrease or increase of electron density in the ionosphere. For this type of rhythmic fading the ionospheric ionization should be just enough for a single reflection to take place between the transmitting and receiving stations; in other words the frequency of the wave should be in the neighbourhood of the maximum usable frequency (M.U.F.) for the relevant ionospheric layer. Rhythmic fading of a different origin has also been reported by Banerjee and Mukerjee (1948) and Banerjee and Singh (1948, 1949) under conditions of high ionospheric ionization. In the present paper an account is given of an investigation on rhythmic or periodic fading which throws some light on the origin of such fading observed with short-wave signals. The investigation was carried out with Calcutta signals of frequency 4,840 kc/s. received at Dacca at a distance of 240 km. from Calcutta in the evening and early night hours of December 1948 and January 1949.

The observed rhythmic or periodic fading patterns were recorded photographically. The paper gives a description of the experimental method and a discussion of the experimental results. In interpreting the results a new viewpoint has been advanced which explains the rhythmic fading observed under certain conditions, when the Appleton–Beynon type of periodic fading was not possible.

§ 2. EXPERIMENTAL ARRANGEMENT

(i) *Receiving Set*

A T.R.F. receiver having one stage of R.F. amplification followed by a grid-leak detector and an A.F. power amplifier (to work a loudspeaker for facilitating tuning) was used with a big frame aerial. The receiver and the aerial tuning condenser were inside a shielded box. A suitable mirror galvanometer kept outside the shielded box was inserted in the anode circuit of the detector valve with a conventional arrangement for balancing out the no-signal anode current.

(ii) *Recording Apparatus*

The recording arrangement was similar to that adopted by Subba Rao and Subramanyam (1942) in their work on atmospherics.

The general disposition of the recording apparatus is shown in Figure 1. With the help of the condensor lenses L_1 and L_2 placed in the paths of the incident and reflected beams to and from the galvanometer mirror M, the light from a point-o-lite was converged to a sharp focus on a bromide paper fixed round the rotating drum D. The drum, which was made of wood, was cylindrical in shape with a circumference of 12 in. and length 10 in. It was enclosed inside a long light-tight box E provided with an opening for admitting light for the photographic recording.

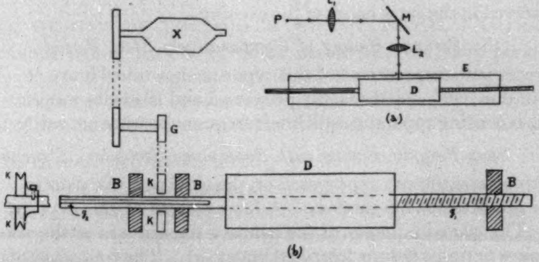


Figure 1 (*a*). General disposition of apparatus:
P—Point-o-lite, M—Galvanometer mirror, L_1, L_2—Convex lenses,
D—Rotating Drum, E—Light-tight enclosure.

Figure 1 (*b*). Working of the drum:
X—Electric motor, G—Reduction gear, B—Bearings, g_1—Spiral groove,
g_2—Linear groove, KK—Pulley fitted with a guide pin.

The brass axle of the drum, which projected on either side through a length of 12 in., was supported on suitable bearings. Along the axle-rod on one side there was a spiral groove of pitch 0·9 cm. and on the other side a linear groove 3 mm. deep. The bearing which supported the axle-rod with the spiral groove on it had an attached pin fitting into the spiral groove. Round the axle-rod with the linear groove on it there was a pulley KK with a guide pin which fitted into the linear groove and could move along it. As the pulley was made to rotate slowly with the help of an electric motor and the reduction gear, the guide pin on the pulley fitting into the linear groove caused the axle and the drum to rotate at the same time. Besides this rotation there was also a linear motion of the drum along its axis due to the spiral groove and the attached bearing. The linear movement was maintained as the guide-pin on the pulley moved along the linear groove.

This arrangement enabled the drum to move laterally through a distance of 0·9 cm. for one complete rotation. Hence the pattern marked out by the galvanometer spot on the bromide paper, when undeviated by the signals, consisted of slightly inclined parallel lines along the length of the paper.

The speed of rotation of the drum was such as to furnish a time scale of 15–40 cm. per minute. For each run of the recording drum, the speed was determined very accurately.

§ 3. EXPERIMENTAL RESULTS

With Calcutta signals of frequency 4,840 kc/s. received at Dacca during the evening and early night hours, the following rhythmic fading patterns were observed :

(i) *Sinuous Fading of Comparatively Quick Period*

Some typical sinuous fading patterns obtained with carrier waves and with fixed tone modulation are shown in Figure 2*. Some records of sinuous patterns which were observed with the usual modulated waves during the time of a broadcast programme are illustrated in Figure 3. The frequency of these patterns ranged roughly between 25 and 60 cycles/minute. The sinuous patterns were usually observed in the early hours of the evening.

(ii) *Periodic Fading of Comparatively Slow Period*

Some representative patterns of this type are shown in Figure 4. The *quasi*-frequency of this type ranged roughly between 2 and 10 cycles/minute. With the speed of the recording apparatus, still lower frequencies were not easily discernible.

(iii) *Slow Periodic Fading with Superposed Secondary Structure*

The secondary structure superposed on the slow periodic pattern was sinuous or periodic and occasionally random. A few patterns of this type are shown in Figure 5. The *quasi*-frequency of the primary pattern was of the same order as that of the slow periodic fading described under (ii). The frequency of the ripples superposed on the slow pattern was of the same order as that of the sinuous fading described under (i). In a few cases the superposed ripples were as rapid as 80–100 cycles/minute (Figure 5).

The slow periodic fading patterns with or without superposed ripples were usually observed in the later hours of the evening.

In some records the superposed ripples were found to have maximum amplitude at each peak of the primary pattern with a gradually decreasing amplitude on either side of the peak. Such secondary structure can be seen in Figure 5 (*b*). Reference should also be made to extremely high frequencies (4 to 12 cycles/second) observed in a few of the sinuous and other patterns (see Figure 3). This may perhaps be identified with the, ' flutter ' phenomena recently reported by Subba Rao and Somayajulu (1949).

That there was little or no *selective* fading was evident from the fact that the carrier waves with no modulation, with fixed modulation and with the usual programme-modulation yielded more or less similar types of fading patterns. Perfectly sinuous patterns were found mostly with no modulation and with constant modulation (Figure 2). Good sine curves were also observed when the broadcast programme was on (Figure 3). In some records there was evidence of slight distortion in the sinuous patterns obtained when the programme was on.

---

* For Figures 2–5 see Plates.

It is worthy of note that *pure* random fading with 4,840 kc/s. signals was less frequently observed. The random variation of signal intensity is not, however, the subject matter of the present paper.

The frequencies or *quasi*-frequencies and other details of the different rhythmic patterns illustrated in the paper are given, figure by figure, in the Table.

Fig.	Date	Nature of signals	Time I.S.T. hr.	min.	sec.	Frequency (cycles/min.) Quick period	Slow period	Remarks
2	16.12.48	Carrier	18	34	20	29·7	—	Sinuous
		Tone	18	35	20	29·7	—	Sinuous
		End of line 5	18	38	40			
3	30.12.48	Programme	18	43		24·5	—	Sinuous
	26.12.48	Programme	19	50		35·4	—	Sinuous
4	26.12.48	Programme	19	37		—	9·6	Slow periodic
5	19.12.48	Programme	18	49		48·8	5·0	Slow periodic with ripples
			to			97·8		
			18	55	38	34·9	—	
5	18.12.48	Programme	20	12	25	79·2	4·7	Slow periodic with ripples
		Finish	20	36	25			

§4. ORIGIN OF THE OBSERVED RHYTHMIC FADING

The sinuous fading pattern of quick period and the slow periodic fading observed with 4,840 kc/s. signals constituted the two main types of rhythmic fading, the origins of which appear to be quite distinct. The origin of the two main types of rhythmic fading observed in the investigation is discussed below.

(i) *Sinuous or Periodic Fading of Magneto-ionic Origin*

The ionospheric data for Calcutta show that between Calcutta and Dacca, the M.U.F. values for the E and F layer transmission of the ordinary waves were about 4 Mc/s. and 13 Mc/s. during the evening hours of the observation months. The ordinary component of the signal of 4·84 Mc/s. therefore invariably passed through the E layer. The extraordinary components were, however, returned from the E layer in the early evening hours when the electron density of the layer was sufficient for the purpose. Since the frequency of the signal was only slightly less than the M.U.F. value for the extraordinary component in the E layer, the condition was favourable for interference to take place between the lower and upper trajectory extraordinary waves in that layer, giving rise to a periodic type of fading, with periodicity depending upon the difference between the signal frequency and the M.U.F. for the extraordinary wave in the E layer. The observed sinuous patterns of 25-60 cycles/minute were probably caused by interference of this type. It is to be noted that since the signal frequency was *considerably* less than the M.U.F. value for the F layer, the interference between the upper and lower trajectory waves (either ordinary or extraordinary) in that layer was out of the question. The interference between the lower ordinary and the lower extraordinary components in the F layer would not also be able to produce a discernible periodicity.

(ii) *Periodic Fading Due to Doppler Effect in Moving Ionosphere**

When the E layer was 'patchy' in the later parts of the evening it was at times possible with 4,840 kc/s. signals to receive the reflected waves from the F region as well as the extraordinary waves from the E layer. When the electron density of the

---

* A brief report of this view-point has been published by one of us (Khastgir 1949).

E layer was too low even for the extraordinary component to get returned, the reception would be due to the F layer transmission only. The electron density* of the F layer in the evenings was usually sufficient for both single and double reflection to take place at the F layer (with 4,840 kc/s. signals the minimum electron densities of the F2 region for single and double reflections should be about $2{\cdot}1 \times 10^5$ and $2{\cdot}8 \times 10^5$ electrons/cm^3 respectively). Thus in the later parts of the evening the downcoming waves usually consisted of the singly and doubly reflected waves from the F2 region.

With a vertical movement of the ionospheric layers in the evening (or early morning) when there is simultaneous reception of the singly and doubly reflected waves from the F2 region, it is evident that they will suffer different amounts of Doppler change of frequency as they proceed towards the receiving point from distinctly different directions. Thus there will be a difference in the frequencies of the singly and doubly reflected waves from the same moving layer. A similar difference in frequency is also expected in the case of simultaneous single reflections from E and F regions. In either case, the two interfering waves of slightly different frequencies would give a resultant beat-note with a progressively increasing amplitude followed by a corresponding decrease in amplitude of periodic character. When the beat-note is received by the receiver, the output after rectification in the detector stage would constitute the envelope of such resultant beat-note with one side wiped out. This would be similar in appearance to a slow rhythmic fading, the periodicity of which would correspond to the difference in the frequencies of the downcoming waves as determined by the Doppler formula.

Let us first consider the case when the singly and doubly reflected waves from the same layer proceed towards the receiving point. It is to be noted that a singly reflected wave suffers a Doppler change of frequency once and a doubly reflected wave twice. If the angles of incidence corresponding to the singly and doubly reflected waves from the same layer moving with velocity v in the vertical direction are denoted by θ_1 and θ_2 respectively, then the difference in frequency Δf between the two interfering waves is given by

$$\frac{\Delta f}{f} = \frac{2v}{\mathbf{c}}(2\cos\theta_2 - \cos\theta_1)$$

when f is the signal frequency and $\mathbf{c}$ the velocity of light. The *quasi*-frequency n of the periodic pattern which is the same as Δf is then obtained from

$$n = \frac{2v}{\lambda}(2\cos\theta_2 - \cos\theta_1) \qquad \ldots\ldots (1)$$

where λ is the wavelength corresponding to the frequency f. Similarly, in the case of singly reflected waves from the two layers moving with the same velocity v in the vertical direction, the *quasi*-frequency of the periodic fading pattern due to the Doppler beat effect is given by

$$n = \frac{2v}{\lambda}(\cos\phi_1 - \cos\theta_1) \qquad \ldots\ldots (2)$$

---

* The ionospheric data from the Calcutta observations under Prof. S. K. Mitra show that on the observation days and time, the electron density of the F2 region ranged from $1{\cdot}55 \times 10^6$ to $2{\cdot}4 \times 10^6$ electrons/cm^3.

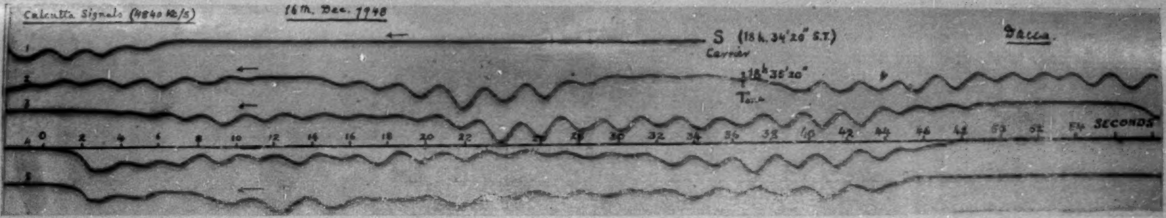


Figure 2. Sinuous patterns. S—Record starts with carrier waves at 18h. 34m. 20s. and tone begins at 18h. 35m. 20s.

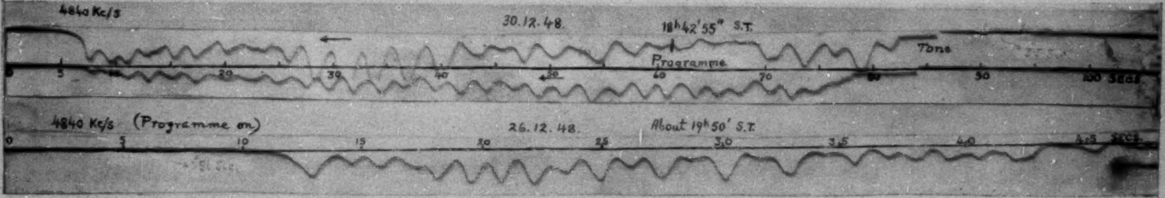


Figure 3. Sinuous patterns with modulated waves.

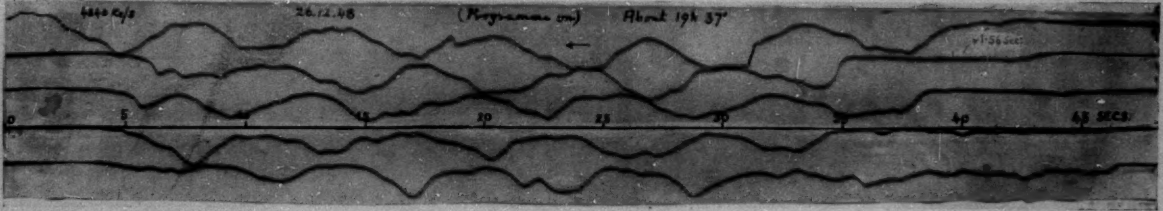


Figure 4. Slow period fading.

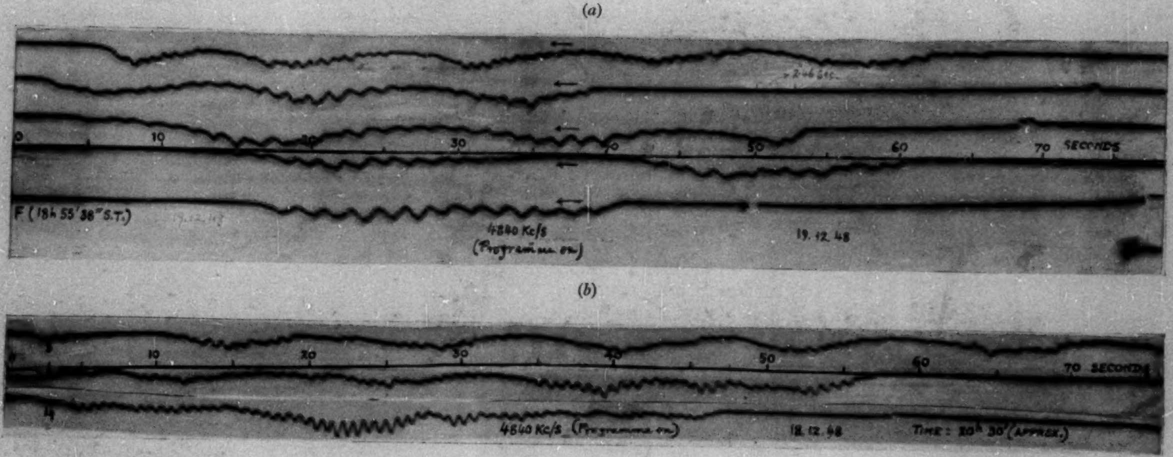


Figure 5. Slow periodic fading with ripples.

where θ_1 and ϕ_1 are the angles of incidence at the E and F layers respectively for simultaneous single reflections.

Considering the signals from Calcutta received at Dacca (distance 240 km.) and taking the E and F2 regions to be at heights of 100 km. and 360 km. respectively, the formula (1) in the case of the singly and doubly reflected waves from the F2 region becomes

$$n \simeq 2v/\lambda \qquad \ldots\ldots(1a)$$

Similarly the formula (2) for the singly reflected waves from both E and F2 regions is reduced to

$$n \simeq 0{\cdot}62\, v/\lambda \qquad \ldots\ldots(2a)$$

If the vertical drift of the ionospheric layer is taken as 4 metres/sec. (about 14 km/hour) during the evening hours, the singly and doubly reflected waves from the F2 region would according to (1*a*) give periodic fading of nearly 8 cycles/minute whereas according to (2*a*) the singly reflected waves from both E and F2 regions would produce periodic fading of nearly 2·5 cycles/minute. The observed slow periodic fading (2-10 cycles/minute) can thus be satisfactorily explained according to the Doppler beat interpretation outlined above.

The observed slow periodic fading with superposed ripples was evidently of a mixed type when the ionospheric conditions were favourable for the simultaneous occurrence of the two types of rhythmic fading already described. Thus the quick periodic fading due to the interference between the upper and lower trajectory extraordinary waves in the E layer was found superposed on the slow periodic pattern caused by the Doppler beat effect between the singly and doubly reflected waves from the F2 region or between the singly reflected waves from the E and F2 regions, when the ionospheric conditions were favourable for such effect. It is evident from the formulae (1 *a*) and (2 *a*) that the 'beating' of the singly reflected waves from the moving E and F2 regions would cause periodic fading of slightly slower period than the 'beating' of the singly and doubly reflected waves from the same moving F2 region.

The vertical velocity of the ionospheric layer or layers as computed from the observed slow periodic patterns according to the Doppler beat view-point is of the order of 10 km/hour (i.e. about 3 metres/sec.) This agrees with the approximate estimate from the ionospheric data for the observation days and time.

Attention should be directed to the work of Banerjee and his colleagues (1948, 1949) who reported observations of periodic fading with short waves under conditions of relatively high ionospheric ionization. They attributed such periodic fading to the interference between the two waves singly and doubly reflected from the same layer, or between the singly reflected waves from both E and F layers, when there was a continuous change in the path difference of the two interfering waves due to the vertical movement of one or both layers. It is to be noted that this view is essentially the same as the Doppler beat interpretation and that both the views are equivalent representations of the same phenomenon. It can be shown that the same expression is obtained for the periodicity of fading according to the two equivalent conceptions*. For oblique incidence, the frequency of fading can be easily calculated from the Doppler beat formulae (1) and (2).

---

* See Khastgir and Das (1950).

§ 5. CONCLUSIONS

The main conclusions of the investigation are summarized below :

(i) Of the rhythmic patterns observed during the evening hours at Dacca with the Calcutta signals of 4,840 kc/s. the sinuous pattern of comparatively quick period is considered to be of magneto-ionic origin as being due to the interference of the upper and lower trajectory extraordinary waves in the E layer, the ordinary waves having passed through the E layer.

(ii) The slow periodic fading observed in the evening was caused by the beat effect between the singly and doubly reflected waves from the F2 region or between the singly reflected waves from the E and F2 regions, the two interfering waves having suffered different amounts of Doppler change of frequency due to the vertical movement of the ionospheric layer or layers. The values of the vertical velocity of the ionospheric layers calculated according to the proposed theory appear to accord well with the observed values.

(iii) The slow periodic fading with superposed ripples was observed when the ionospheric conditions were favourable for the simultaneous occurrence of the magneto-ionic type of sinuous fading and the Doppler beat type of slow periodic fading.

(iv) There was evidence of extremely high frequencies (4–12 cycles/second) in a few patterns of periodic fading, the origin of which is unknown.

ACKNOWLEDGMENT

Our sincere thanks are due to Prof. S. K. Mitra, for kindly supplying us with the relevant ionospheric data for Calcutta.

REFERENCES

Appleton, E. V., and Beynon, W. J. G., 1947, *Proc. Phys. Soc.*, **58**, 59.
Banerjee, S. S., and Mukerjee, G. C., 1948, *Phil. Mag.*, **39**, 697.
Banerjee, S. S., and Singh, R. N., 1948, *Indian J. Phys.*, **22**, 413.; 1949, *Science and Culture*, **14**, 293.
Khastgir, S. R., 1949, *Science and Culture*, **15**, No. 3 (Sept.).
Khastgir, S. R., and Das, P. M., 1950, *Science and Culture*, **15**, 445.
Subba Rao, N. S., and Somayajulu, Y. V., 1949, *Nature, Lond.*, **163**, 442.
Subba Rao, N. S., and Subramanyam, 1942, *Andhra University Journal*.

Magnetic Hysteresis of Igneous Rocks

By J. McG. BRUCKSHAW and B. S. RAO
Imperial College, London, S.W.7

Communicated by A. O. Rankine; *MS. received* 19*th May* 1950

ABSTRACT. A rapid method has been developed for the investigation of the magnetic properties of igneous rocks. The specimens examined show the normal properties to be expected from a mixture of magnetically different constituents. In particular, high coercivities have been found. The significance of the results in relation to natural permanent magnetism in rocks is discussed.

§ 1. INTRODUCTION

It has been shown that a number of igneous bodies, such as the tholeiite dykes of the North of England (Bruckshaw and Robertson 1949), the Pilansberg dyke (Gelletich 1937) etc., possess permanent magnetism whose direction differs materially from the present direction of the earth's magnetic field. The simplest explanation of this phenomenon is that the ambient magnetic field existing at the time when the bodies cooled through the Curie points of their magnetic constituents, had the direction of the observed permanent magnetism. In the case of the tholeiite dykes, this requires at least a local reversal of the field. Since this does not conform with present theories of the origin of the geomagnetic field, it was necessary to examine the magnetic properties of the rocks to ensure that the adverse magnetism could not arise from some anomalous behaviour of the material. Further, the age of the intrusions varies up to several hundred million years, and to retain adverse magnetism over such a long period a large coercivity is necessary. The present investigation is concerned with the examination of the magnetic hysteresis and a measurement of the coercivity of igneous rocks.

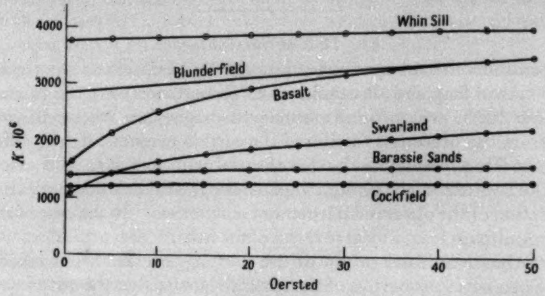
§ 2. PRELIMINARY INVESTIGATIONS

Initially, some observations were made of the variation of the mean volume susceptibility K with different amplitudes of the applied field. For this purpose a method (Bruckshaw and Robertson 1948), previously employed to determine rock susceptibilities in weak fields of about 0·5 oersted, was adapted to measurements up to 50 oersted. In this method an alternating magnetic field is set up by a Helmholtz coil system. In the spatially uniform region at the centre of the system a rock specimen, in the form of a 2 cm. cube, is introduced and, under the action of the applied field, it becomes an alternating magnet which approximates closely to a dipole. When the influence of the applied field is compensated, the E.M.F. set up in an adjacent coil is thus a measure of the induced intensity of magnetization. From this and the known field the mean susceptibility can be determined.

The specimens were samples of tholeiite from the dykes of North England, of dolerite from the Whin Sill and basalts from flows in Nigeria. The Whin Sill, 2×10^8 years old, possesses discordant magnetism (Hallimond and Butler 1949) while the Nigerian basalts also exhibit peculiar magnetism. Several of each type were examined, and typical results of the susceptibility changes up to peak fields of 50 oersted are shown in Figure 1. In the majority of cases the mean susceptibility varied linearly with the peak field, a result consistent with the observations of

Hallimond and Herroun (1943) and the overall change in magnitude was of the order of 10 per cent. The Nigerian basalts, and the samples of tholeiite collected at Swarland, however, gave curves which were not linear and the increase in susceptibility over the whole range of field strengths was large in comparison with the others, in one case as much as 175 per cent. Even in these cases there was no sign of attaining a maximum susceptibility, a feature exhibited by all ferromagnetics.

The absolute values of susceptibility are of little significance, since they merely reflect the content of magnetic material, usually assumed to be magnetite, in the rock. For the tholeiite dykes this content ranged from 3·5 to 4·5 per cent by weight, and for the dolerite about 4 per cent, both values corresponding to about 2 per cent by volume. Slichter (1929) and Nettleton and Elkins (1944) have suggested that the susceptibility of a rock containing disseminated magnetite is given closely by the equation $K=0{\cdot}3P$, where P is the fraction by volume of magnetite On the basis of this formula, the present measurements correspond to a magnetite content varying from 0·04 to 0·94 per cent for the tholeiites and from 1·12 to 1·24 per cent for the dolerites. Thus it would appear that the proposed formula is not applicable and it will be shown later that it is impossible for such a simple formula to predict accurate values.



Figue 1.

§ 3. THE MEASUREMENT OF HYSTERESIS

Many observers have investigated the magnetic hysteresis of rocks, usually using the well-known magnetometer or ballistic methods. The main difficulty is the small intensities of magnetization, usually about 1 gauss in fields of several hundred oersted, in comparison with those normally encountered in ferromagnetic materials. With the increased sensitivity difficulties arise from instability, particularly in towns. In the present method, the properties are displayed by means of a cathode-ray oscillograph, and the principle of the method is illustrated in Figure 2. An insulating former, of the shape shown, carried five coils of which A_1, A_2 were used to produce an alternating magnetic field over the specimen S, and the three coils B_1, B_2 and B_3 to detect the effect of the specimen. These last coils were joined so that the E.M.F's generated in B_1 and B_3 opposed that in B_2. In the absence of the specimen, the number of turns on B_1 and B_3 was adjusted so that the resultant E. M. F. from the three coils was small. To retain a fine control over this adjustment during the experiments, another coil B_3', of a few turns, was wound over B_3 and joined to a high resistance potentiometer P, from which a small variable potential difference, of either sign, could be injected into the circuit to obtain the best balance.

When the specimen was introduced, the induction in the volume occupied by it changed from H to $(H+4\pi I)$ where I is the magnetic moment per unit volume. Since the influence of the time variation in H had been compensated, I only was effective. Because of the different distances involved, the E.M.F. in B_2 due to the changing value of I was greater than the sum of the corresponding E.M.F's in B_1 and B_3, and a resultant E.M.F. proportional to dI/dt was set up across the combination. This E.M.F., which was applied to the Y-sweep of a cathode-ray oscillograph, after suitable amplification, was proportional to $K\,dH/dt$ where $K(=dI/dH)$ is the incremental volume susceptibility given by the slope of the (I,H) graph. The potential drop across a resistance R, in series with the exciting coils A_1, A_2, was applied to the X-sweep and, since the field is proportional to the current, yielded a deflection which was a measure of the instantaneous field. The pattern on the cathode-ray oscillograph was thus a graph showing the variation of $K\,dH/dt$ with H. With a sinusoidal field $H=H_0 \sin \omega t$, the instantaneous vertical deflection is $KH_0 \cos \omega t$ and, for a constant susceptibility, the pattern would be an ellipse.

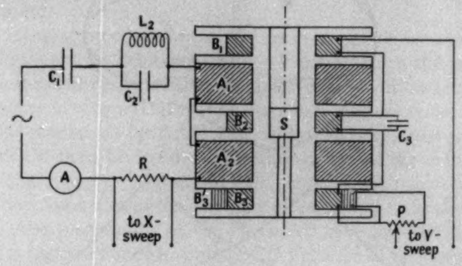


Figure 2.

§4. THE COIL SYSTEM AND THE AUXILIARY CIRCUIT

Details of the coil system are given in Table 1. The inner radius of the coils was fixed by the size of the rock specimen (a 2 cm. cube) and the strength of the insulating former, while the nearer sides of the exciting coil were controlled by the width of the measuring coil B_2 and the insulating partitions. With these two dimensions fixed, the radial width and length were calculated to give the maximum field per unit current for the length of wire employed. Small departures from the optimum dimensions were not important. Close agreement was obtained between the estimated field and that measured by two different methods. The peak field was 246 oersted per r.m.s. amp. and, with a standard current of 3·05 amp., a

Table 1

Coil	Inner radius (cm.)	Outer radius (cm.)	Length (cm.)	No. of turns	Wire (S.W.G.)
A_1A_2	4·2	6·3	2·9	960	20 En. & S.C.
B_3	4·2	—	1·4	1940	40 D.C.C.
B_1B_2	4·2	—	1·4	1850 (approx.)	40 D.C.C.

field of 750 oersted was obtained. The uniformity of the field over the space occupied by the specimen was examined and found to be constant to within 2 per cent.

The power supply was a generator, of 532 c/s., driven by a synchronous motor. To obtain the necessary current from the 110 v. supplies, the circuit was tuned by the condenser C_1. A third harmonic proved troublesome and its influence was substantially reduced by a parallel tuning circuit L_2C_2 adjusted to the harmonic. In balancing the coils B_1, B_2, B_3, their self-capacities gave small phase displacements between the potential differences across their terminals, and it was necessary to place a small variable condenser C_3 across one coil to obtain a suitable balance. The adjustments were correct only for one frequency and a residual pattern was left due to imperfectly balanced harmonics. As this pattern intersected itself it was found convenient to open it out by adjusting the potentiometer P. Frequently, to accommodate large deflections due to the specimen, the initial trace was adjusted so that the specimen caused the trace to collapse and open out in the opposite direction.

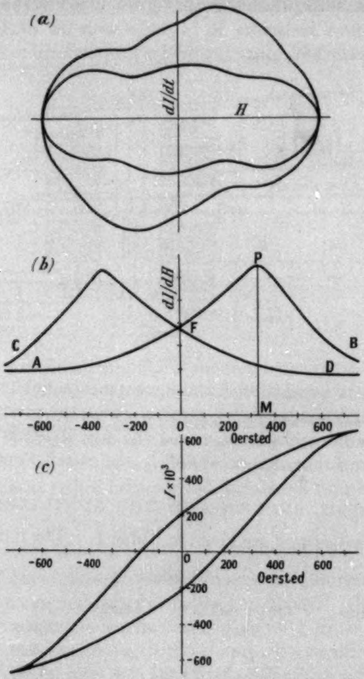


Figure 3.

A photographic record was made for each specimen, the record consisting of the trace without the specimen, the trace with the specimen in position and the two axes, the last being obtained by disconnecting first the X-sweep and then the Y-sweep. A reproduction of a typical record is shown in Figure 3 (*a*).

§ 5. ANALYSIS OF THE RECORDS

Subtracting the deflections obtained with and without the specimen gave values proportional to $KH_0 \cos \omega t$ for various values of the applied field. The latter value was known in oersted since the maximum X-deflection corresponded to the peak field of 750 oersted. As the combination of amplifier and cathode-ray oscillograph did not give deflections strictly proportional to the input, corrections were necessary to take this into account. For any given value of H, the value $\omega t = \sin^{-1}(H/H_0)$ could be determined and hence the Y-deflections converted into values proportional to K. These values were plotted, yielding curves of the type shown in Figure 3 (*b*), in which the curve AB corresponds to the field increasing (upper part of the original record) and DC to the field decreasing (lower part of record). As would be expected, the latter is the mirror image of the former in the Y-axis. It should be noted that the susceptibilities at the maximum and minimum fields are extrapolated, since both $\cos \omega t$ and the deflection are zero here and the ratio indeterminate.

Numerical integration of the $(dI/dH, H)$ curves yielded the shape of the (I, H) loop with the intensity in arbitrary units. Further, since the integration must commence at some arbitrarily selected zero, the position of the curve in relation to the axis of intensity of magnetization was unknown. In the cyclic condition, it was assumed that the intensities for fields $+H_0$ and $-H_0$ were equal and opposite, and the line $I = 0$ could then be fixed. The intersection of the curve with this line permitted the coercivity to be determined. Integration round a complete cycle was used to obtain some idea of the accuracy of the results. In the example shown in Figure 3 (*b*), the integration of the curve AB gave an overall change of 45·1 arbitrary units of intensity and the branch CD a change of 44·1 units. The former half cycle yielded a coercivity of 183 oersted and the latter 180 oersted. Discrepancies of a similar magnitude were found in other cases, and the mean coercivity quoted here is probably accurate to 2 or 3 oersted. Some idea of the magnitudes of the susceptibility and of the intensities of magnetization can be obtained by assigning to OF of Figure 3 (*b*) the susceptibility observed in the preliminary experiments using a small field. The intensities shown in Figure 3 (*c*) were derived on this basis. This substitution, obviously, is not strictly correct, but the error involved cannot be serious. Indeed, for the present purpose the absolute values, reflecting the concentrations of magnetic materials, are of little significance. A direct calibration of the system is possible, for by using the equivalent of a current carrying loop and a magnetic shell, a coil wound with n turns per unit length corresponds to a volume intensity of magnetization of ni, where i is the current in E.M.U.

§ 6. EXPERIMENTAL RESULTS

The rock samples examined in the preliminary tests were also examined in the stronger fields. From the measurements two values characteristic of the specimen, could be determined, the value H_m of the field corresponding to the maximum susceptibility and H_c, the coercivity. The ratio of the maximum susceptibility to the value at $H = 0$, i.e. PM/OF in Figure 3 (*b*), should also be characteristic but as this varied little, from about 1·5 to 1·7, it was not used. The results for most rocks are shown in Table 2. Examination of Figure 3 (*b*), in which the susceptibility does not fall to zero at high fields, and which also exhibits a discontinuity on reversal, shows that the specimen did not attain saturation*. This was true in all cases, and the coercivities given here are underestimates of the true coercivities. There is a corresponding uncertainty in H_m. Nevertheless, as each rock was examined under standard conditions a comparison of the results is valuable.

---

* On the basis of disseminated spheres of magnetite an estimated field of about 2,200 oersted would be required for saturation.

The results of Table 2 refer to those specimens which gave simple curves. The tholeiite specimens from the Swarland exposure and the Nigerian basalts, however, gave anomalous results, the former exhibiting two maxima on the susceptibility curve and the latter three maxima. The original cathode-ray oscillograph pattern for a sample from the Swarland exposure is reproduced in Figure 4. The results for the anomalous specimens are summarized in Table 3.

Table 2

Specimen	H_m (oersted)	H_c (oersted)	$K_0 \times 10^6$
THOLEIITES			
Aklington Dyke			
Abington exposure	338	165	1920
	375	180	1960
	263	120	2640
Lugton exposure	255	113	1920
Cockfield exposure	319	195	980
Cleveland Dyke			
Blunderfield exposure	223	113	1290
	338	182	1050
Hazel Cottage exposure	263	165	2000
	338	180	2000
Barassie Sands exposure	450	270	1400
DOLERITE			
Whin Sill at Cross Fell	300	188	3520
	300	210	3400

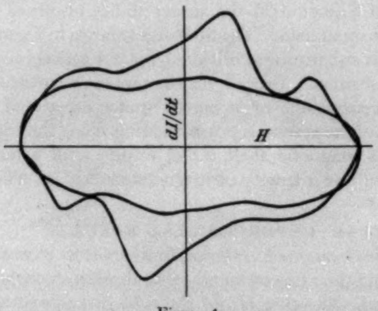


Figure 4.

§7. DISCUSSION OF RESULTS

The most obvious interpretation of the multiple maxima on the susceptibility curve is the existence within the rock of two or three magnetic materials with distinct properties. This is consistent with observations, made at the Imperial College and elsewhere, on the variation of susceptibility with temperature in which a specimen may show several Curie temperatures distinct from the generally accepted value of 580° c. for magnetite. For specimens with more than one component, the observed values of H_m and H_c will not have any precise significance. Since the peak due to one component is superimposed on a rising or falling background due to the other materials, a displacement of the peak will occur. The coercivity, on the other hand, will be intermediate between the greatest and least coercivities of the constituents. During the demagnetization of the specimen, when the field attains the smallest coercivity, some components will still retain their magnetism while, when the greatest coercivity is reached by the field, other components will possess magnetism in the direction of the field. The condition that the average intensity of magnetization is zero will be attained between these two extremes, at a field strength dependent on the relative concentrations and the intrinsic properties of the individual constituents.

The results in Table 2 in which the specimens appear to show only one magnetic component, demonstrate that there is little constancy in either H_m or H_c, the former ranging from 223 to 450 oersted and the latter from 113 to 270 oersted. Even in samples from the same exposure, considerable variations are apparent, as illustrated by the three collected at Abington.

Assuming that the properties of the magnetic constituents are similar to those of magnetite, the order of magnitude of H_m is to be expected. Specimens of natural magnetite give maximum susceptibilities at fields lying between 20 and 50 oersted, when they possess intensities of magnetization of the order of 50 gauss. Assuming in the rock spherical particles, with a demagnetization factor of $4\pi/3$, an external field of 260 oersted is necessary to produce an internal field of 50 oersted. Fluctuations in the mean demagnetization factor for the particles would account for some of the variation in the value of H_m. This, however, would appear to account for only a small part of the observed range for, when $I=0$, the internal and external fields are equal. Thus, if all of the disseminated particles had the same magnetic properties, the rocks should possess the same coercivity. There is thus clear evidence that a wide variety of magnetic properties is involved, even in those specimens which, by virtue of a single peak, appear to contain one constituent only. Such a variety may be attributed to different impurities, particle size and heat treatment. The last factor is necessarily connected with the second. Basalt flows in general cool rapidly and show a fine grained, and even vitreous, structure. There is also the possibility of reheating by subsequent flows over them. Sills and dykes, on the other hand, will cool more slowly, but the reduction of temperature will be faster in the parts in contact with the country rock than in those near the centre. Such a differentiation of cooling may well play a part in accounting for the variations in samples from the same exposure.

Table 3

Specimen	H_m (oersted)	H_c (oersted)	$K_0 \times 10^6$
Tholeiites Swarland exposure	173, 563	98	720
	280, 660	189	500
Basalts Nigeria	137, 516, 720		600
	208, 572, 675		900
	150, 525, 675		640

Although it seems probable that there is a continuous range of properties arising from different factors, it must not be assumed that each specimen in Table 2 contains only one type of material because it exhibits one maximum on the susceptibility curve. This curve will only resolve the influence of the individual constituents if the peaks are separated by an interval greater than a critical amount. At smaller intervals the two will merge into a single broad peak. The critical interval is obviously related to the relative magnitudes of the peaks and their shape.

It is apparent that in these conditions the simple formula, suggested by Slichter for the susceptibility in weak fields in terms of magnetic content, is untenable. If the range of properties is due to a finite number of constitutents, it might be possible to replace it by a sum of similar terms, each relating to a single constituent. Even this seems improbable and the formula must be regarded as an over-simplification.

The coercivity values which have been determined are, for reasons given, underestimates of the greatest coercivity involved. Even so, they are high in comparison with the values for iron, cobalt, nickel etc. for which the values range from 1 to 12 oersted. The present values are comparable with those of modern magnet steels. They are similar to those obtained by other investigators, such as Puzicha (1930) who obtained values from 5 to 226 oersted. If the permanent magnetism of the tholeiite dykes was, as suggested, acquired shortly after their injection, it has been retained for 20 to 30 million years against a demagnetizing external field for a large part of this period. The Whin Sill, on the same assumptions, has retained its magnetism for several hundred million years, as also must the Pilansberg Dykes of S. Africa. It is difficult to convert the coercivities as measured into loss of magnetic moment with time. Many factors may assist in demagnetization such as earth tremors, stresses due to earth movements, temperature changes etc. Nevertheless, the high coercivities make reasonable the assumption that, if the earth's field has not changed appreciably in magnitude, some of the original intensity of magnetization would be retained.

Finally, it is obvious that the magnetic behaviour of these rocks is normal and it does not display any anomaly which might account for the observed adverse polarization of many igneous bodies.

ACKNOWLEDGMENT

The authors wish to record their thanks to the Anglo-Iranian Oil Company whose financial assistance made this investigation possible.

REFERENCES

Bruckshaw, J. M., and Robertson, E. I., 1948, *J. Sci. Instrum.*, **25**, 444; 1949, *Mon. Not. Roy. Astr. Soc., Geophys. Suppl.*, **5**, 308.
Gelletich, H., 1937, *Beitr. Geophys.*, **6**, 337.
Hallimond, A. F., and Butler, A. J., 1949, *Mining Mag.*, **80**, 265.
Hallimond, A. F., and Herroun, E. F., 1943, *Proc. Phys. Soc.*, **55**, 214.
Nettleton, L. L., and Elkins, T. A., 1944, *Geophysics*, **9**, 60.
Puzicha, K., 1930, *Z. prac. Geol.*, **38**, 161, 184.
Slichter, L. B., 1929, *Geophysical Prospecting, Amer. Inst. Mech. Engng.*, New York, 238.

The Michelson Interferometer at Millimetre Wavelengths

By W. CULSHAW

Telecommunications Research Establishment, Ministry of Supply, Great Malvern, Worcs.

Communicated by R. A. Smith; *MS. received 27th April* 1950

ABSTRACT. The design and operation of an interferometer of the Michelson type at a wavelength of 12·5 mm. is discussed.

The required frequency stabilization of the 12·5 mm. source is achieved by using a high Q cavity as an R.F. discriminator.

Wavelength measurements have been made with various spacings of the interferometer, the measured wavelength increasing as the spacing is reduced. Results indicate that the interferometer gives more accurate measurements of the wavelength when operated well in the Fraunhofer region of diffraction.

The operating frequency has been measured using a calibrated frequency meter, and hence the velocity of electromagnetic waves deduced. The value obtained for this velocity agrees, within the accuracy attempted with the present instrument, viz. one part in 10^4, with the generally accepted value for this velocity. The possibilities of increasing the accuracy of its determination with the interferometer are discussed.

Measurements of the dielectric constants of low-loss materials by means of the interferometer have been made. Results obtained using ordinary commercial sheets of materials agree within a few per cent with values obtained otherwise.

The use of the interferometer as a substandard of length is also discussed.

§ 1. INTRODUCTION

In order to ensure that propagation occurs only in the dominant or fundamental mode, the lateral dimensions of conventional transmission systems, such as waveguides, are made smaller than the operating wavelength. Consequently for operation at millimetre wavelengths the dimensions of waveguides are small and the attenuation is large. For a similar reason the dimensions of components such as cavity wavemeters are small in the millimetre wavelength region, and the resulting increased attenuation adversely affects their discrimination. In addition there is a decrease in the tolerance permissible in the manufacture of such components. It thus becomes apparent that for these small wavelengths alternative forms of transmission systems and measuring instruments are desirable, whose lateral dimensions are not limited by the wavelength. In this connection it is natural to consider the application of optical techniques to millimetre wavelengths.

The principle of any optical instrument can be applied to millimetre wavelengths. It cannot be assumed, however, that the resulting millimetre wave instrument will perform its function as efficiently as its optical counterpart; the effect of the difference in wavelength must be considered, both on the technique of the particular measurement and on the physical properties of materials used in the construction of the instruments. The most important limitation in this connection is the diffraction which occurs at millimetre wavelengths with radiating apertures of practical dimensions. With the generators available at present the aperture dimensions are restricted to about 15 wavelengths. The beams obtained are consequently divergent, the constant phase fronts being by no means plane, but dependent on the distance from the radiating aperture.

The problems of obtaining an efficient transmission system and precise measuring techniques at millimetre wavelengths are simpler, as far as application of optical techniques is concerned, the smaller the wavelength. As in many optical instruments, for instance the optical telescope, there is also the possibility of screening the system, but again it is necessary to consider the effect of the difference in wavelength. In the millimetre wavelength region with practical tube sizes the wavelength in the tube depends on its dimensions and the free space wavelength is not obtained. In addition there is the possibility of exciting other modes of propagation in such tubes with a consequent loss of energy and discrimination.

In order to study some of the problems in this field the principles used in the Michelson interferometer have been adapted for operation at a wavelength of 12·5 mm. and an instrument has been constructed. This paper describes the various factors involved in the design and operation of the instrument together with the results of measurements made with it.

§2. DESIGN CONSIDERATIONS

2.1. *General Principles*

The optical form of the Michelson interferometer is well known (Wood 1934). Light from an extended monochromatic source is divided by a ' half-silvered ' glass plate and caused to travel different paths at right angles to each other. The beams are reflected by mirrors and combine at the observer. Due to the multiple reflections in the interferometer, the mirrors can be regarded as being on the same straight line and separated by a distance t, where t is the difference of the distances of the mirrors from the ' half-silvered ' glass plate.

If the mirrors are parallel then we have a maximum intensity when

$$2\mu t \cos\theta = n\lambda, \qquad \ldots\ldots(1)$$

where λ is the free space wavelength, μ is the refractive index, and $n=0, \pm 1, \pm 2$ etc. We have a minimum intensity when

$$2\mu t \cos\theta = (2n+1)\lambda/2; \qquad \ldots\ldots(2)$$

$n=0, \pm 1, \pm 2$ etc.

On focusing the eye at infinity a system of circular fringes is seen, since θ the angle of incidence varies when an extended source is used. In general, therefore, complete darkness is not obtained, the fringe system, or energy distribution across the aperture merely changing with distance t. The intensity variation of the fringes is of the form of $\cos^2\phi$, the maxima and minima diffusing into each other.

For operation in the millimetre wave region transmitting and receiving apertures are used in place of the extended light source and the ' eye '. Since energy is received over the area of the receiving aperture, this necessitates the use of essentially parallel beams in the interferometer in order to obtain sharp maxima and minima (i.e. it is desirable to work with a single fringe). It is also essential to consider diffraction effects in relation to the size of apertures and the spacing of the component parts of the interferometer.

A difficulty encountered in the theory of diffraction of electromagnetic waves is the determination of the field in the radiating aperture consistent with the boundary conditions. The most familiar theory is that of Huyghens–Kirchhoff (Stratton 1941) which assumes that the field in the aperture is that which would exist if the conducting part of the screen were not present.

This theory gives good approximations for apertures whose dimensions are large compared with the wavelength of the radiation, as is usually the case in optics where the Huyghens–Kirchhoff formula forms the whole basis of diffraction theory. The more familiar Fraunhofer and Fresnel types of diffraction follow from this general formulation.

At millimetre wavelengths the aperture is not always large compared with the wavelength, and some care is necessary in the application of the above formula. However, it has a wide application in the field of aerial theory with aperture dimensions of the order of 10 wavelengths, and there is reason to suppose that it will give reasonable practical results for both relatively near and distant diffraction patterns when the dimensions of the aperture are of this magnitude.

A consideration of the diffraction patterns at various distances from a long slit some 15 wavelengths wide at millimetre wavelengths leads to the following conclusions, the slit being supposed uniformly illuminated by a plane wave front.

(i) If λ is the free space wavelength, a the width of the slit, and z the distance along the normal to the slit, then when $\lambda z > a^2$ Fraunhofer diffraction methods can be employed. These give the diffracted field at relatively large distances from the radiating aperture. For this region, near the z axis through the centre of the slit, the constant phase fronts are essentially spherical, the curvature decreasing as z increases, and the amplitude of the diffracted field is inversely proportional to z.

(ii) For $\lambda z < a^2$, but not below a certain value which depends on the obliquity factor and the inverse variation with distance of the amplitude of the Huyghens' sources, Fresnel diffraction methods are used. In this region, on planes perpendicular to the z axis, the amplitude of the diffracted field fluctuates through maxima and minima, decreasing rapidly as the geometrical shadow is approached but extending beyond it. The shapes of the amplitude and phase diagrams of the diffracted field vary rapidly with distance z, the constant phase fronts being by no means plane.

(iii) The distance z_0 given by $\lambda z_0 = a^2$ may be called the transition distance between the regions of Fraunhofer and Fresnel types of diffraction for the particular case here considered. The criterion should be applicable to rectangular and circular apertures provided the appropriate width is substituted, the apertures again being uniformly illuminated by a plane wave front. The criterion gives the minimum distance at which the usual polar diagram of such radiating apertures may be measured, and for a given wavelength is directly proportional to the gain of the aperture.

Ideally the interferometer requires plane wave trains for its operation in the millimetre wave region, but since this is not practicable it is more likely to function most efficiently in a region of uniform diffraction such as in the Fraunhofer region, rather than in the region of Fresnel diffraction where violent changes occur in the amplitude and phase of the diffracted field. In order to preserve a balance between the two beams as mirror 1 is moved over its traverse, and also to obtain adequate energy transfer through the interferometer, the divergence of the radiated energy must be small. The practical size of the apertures used is thus a compromise between the beam divergence and the distance from the radiating aperture to the Fraunhofer region of diffraction, since the larger we make this aperture the greater this distance becomes. The transmitting and receiving apertures in the present interferometer are 6 in. × 6 in., the linear dimensions thus being about 12 wavelengths at a wavelength of 12·5 mm. An aperture of these dimensions supposed uniformly illuminated has an angle of 10° between the first zeros either side of the main lobe of the Fraunhofer diffraction pattern, and the distance z_0 from the aperture to the Fraunhofer region as given by the above considerations is about 180 cm. This means that for operation in the Fraunhofer region the two path lengths in the interferometer between the transmitting and receiving apertures should both be greater than 2 metres.

2.2. *Transmitting and Receiving Apertures*

The transmitter consists of a reflection klystron valve generating radiation of wavelength about 12·5 mm. Power from this is fed into a rectangular waveguide of sectional dimensions 0·170 in. × 0·420 in., and the waveguide is then gradually enlarged in the form of a pyramid to form a horn with aperture 6 in. × 6 in. In order to approach the condition of uniform illumination, and thus to reduce the divergence of the transmitted beam, the horn is terminated by a metal lens of the appropriate focal length. This lens transforms a spherical wave diverging from the point of intersection of the sides of the pyramid into a plane phase front at the radiating aperture.

The receiving aperture consists of a similar horn which is tapered to the size of the waveguide, the received signal being fed into a millimetre-wave crystal detector. The receiving horn is also phase corrected by a metal lens in order to increase its effective receiving area. The increase in size from the waveguide of sectional dimensions 0·170 in. × 0·420 in. to the 6 in. × 6 in. aperture takes place over a length of 50 cm., and the metal lenses employed have focal lengths of 52·5 cm. The design of these 'metal lenses' is now well known (Kock 1946). The fact that the phase velocity of electromagnetic waves in a waveguide is greater than in free space is utilized to simulate a medium of the appropriate refractive index. The profile of the lens is then calculated using geometrical optics. The metal lenses used have a refractive index of 0·6 and hence a plate spacing of 7·81 mm., and they have proved very effective in obtaining an efficient transfer of energy through the interferometer. Figure 1 shows a schematic diagram of the millimetre-wave interferometer and Figure 2 is a photograph of the equipment used.

2.3. *Beam Divider*

In the optical form of the Michelson interferometer a 'half-silvered' glass plate is used to produce the two beams required; ideally this 'beam divider' sends 50% of the incident energy along each path. In view of the fact that silver is almost a perfect reflector at millimetre wavelengths, this technique did not seem practicable, and other forms of 'beam divider' have been used. Experiments were made using various types such as wire and slot gratings, designed to reflect 50% of the energy incident at 45°. These gratings have operated quite well in the interferometer, but were superseded by a more efficient and adjustable reflector which consisted of two quarter-wave plates of dielectric with an adjustable air gap between them as shown in Figure 1. The plates are quarter-wave thick for 45° incidence and their separation is adjusted to give a power reflection coefficient of 50%. This technique has proved satisfactory using polystyrene plates of dielectric constant about 2·5. The adjustable dielectric beam divider has been useful in another way. If the wave fronts in the interferometer were plane, the balance at a minimum would not depend on the reflecting properties of the 'beam divider', since there is a transmission and a reflection for each path. Thus the balance of the two beams would be unaffected

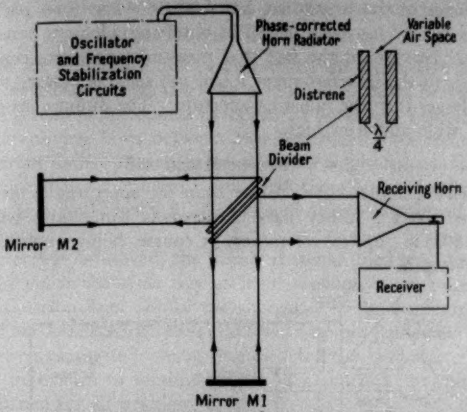

Oscillator and Frequency Stabilization Circuits
Phase-corrected Horn Radiator
Variable Air Space
Distrene
$\frac{\lambda}{4}$
Beam Divider
Receiving Horn
Mirror M2
Receiver
Mirror M1

Figure 1. Millimetre-wave Michelson interferometer.

by the value of the reflection coefficient, although the sensitivity and energy transfer from transmitter to receiver would be greatest for a power reflection coefficient of 50%. However, in the present interferometer the beams are divergent, and any asymmetry in the two paths will destroy the balance, but since the reflectivity of the dielectric ‘ beam divider ’ can be adjusted to compensate

Figure 2. Millimetre-wave Michelson interferometer.

for this, a balance of the beams can be obtained when their path lengths are different. This facility has been useful in obtaining adequate sensitivity in the interferometer. In view of the fact that measurements were contemplated at various spacings of the interferometer it did not seem necessary to use a beam divider much larger than the radiating aperture. The dimensions of the divider used are 9 in. wide and 8 in. high.

2.4. *Mirrors M*1 *and M*2

The mirrors are made slightly larger than the apertures in order to reduce possible 'edge effects'. They have dimensions 8 in. × 8 in. with a surface tolerance of 0·0005 in., optical tolerances, of course, being unnecessary at these wavelengths. The mirrors are machined from brass which is then silver-plated. One of the mirrors, M1, is mounted on a V-slide and is movable by means of a lead screw of 2 mm. pitch over a distance of 50 cm., with a reading accuracy better than 0·0005 cm.

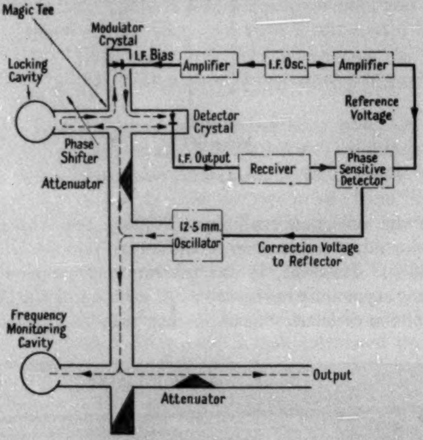

Magic Tee
Modulator Crystal
I.F. Bias
Amplifier
I.F. Osc.
Amplifier
Locking Cavity
Reference Voltage
Detector Crystal
Phase Shifter
I.F. Output
Receiver
Phase Sensitive Detector
Attenuator
12·5 mm. Oscillator
Correction Voltage to Reflector
Frequency Monitoring Cavity
Output
Attenuator

Figure 3. Block diagram of frequency stabilization system used on interferometer.

2.5. *Frequency Stabilization*

From observations made using a cavity wavemeter to measure the frequency stability of the 12·5 mm. source, it became clear that if precise measurements of wavelength were to be made by means of the interferometer, then some frequency stabilization of the source was necessary. A high degree of frequency stability has been achieved using the Pound Frequency Stabilizer (Pound 1946), in which the 12·5 mm. source is effectively locked to a cavity of high Q. The system employed in the interferometer is shown in Figure 3, an additional cavity with a Q value of some 18,000 being provided for frequency monitoring. The frequency of the source is locked to the frequency controlling cavity, which is then tuned to the frequency monitoring cavity so that the operating frequency is kept constant.

The frequency stability of the present system is about one part in 10^5 for c.w. operation, and could be improved by suitable thermostatic control.

§ 3. OPERATION OF THE INTERFEROMETER

3.1. *General Considerations*

Apart from the difficulties due to diffraction, the Michelson interferometer at millimetre wavelengths operates with essentially parallel beams, so that the circular fringe system of the optical model is not obtained. Instead, interference occurs at the receiving horn between two sinusoidal wave trains which have traversed different paths, producing essentially a single fringe. When the path difference is such that a minimum is obtained, there is an effective short circuit in the receiver arm, and ideally all the radiated energy is reflected back into the transmitting horn.

When mirror M1 is moved, the received signal fluctuates through maxima and minima in much the same way as in a standing-wave pattern on a short-circuited transmission line. Since the sharpness of the minimum depends on the equality of two beams which have traversed different paths, this sharpness will vary as M1 moves along its traverse, due to diffraction and scattering effects, so that the high maximum to minimum ratio of a short-circuit transmission line cannot be obtained for all positions of M1.

As might be expected, the setting-up of the millimetre-wave interferometer is not as critical as the optical model. The slide of mirror M1 is first adjusted by means of the screws provided so that M1 moves along the axis of the transmitting horn, the mirror M1 already being fixed in the plane perpendicular to its slide. This is best done visually, or using the beam divider to reflect the energy returned from mirror M1 into the receiving horn mounted at right angles to this axis, the slide is adjusted for maximum received signal when mirror M1 is at either end of its traverse. Mirror M2 is now placed in position and adjusted for a maximum received signal, both mirrors being at the same distance from the beam divider. Adjustment of the beam-divider spacing for maximum received signal is also made at this stage. Finally, slight adjustments of the mirrors and beam divider are made to give adequate sharpness to the minima.

3.2. *Measurement of Wavelength and Frequency*

The apparent wavelength of the radiation as measured with the millimetre wave interferometer is obtained by noting the readings of minima which occur as M1 is moved over its traverse. The distance between adjacent minima should be half a wavelength, so that by observing the distance over which M1 moves for a given number of minima the apparent wavelength at the particular spacing of the interferometer can be determined. Once the apparent wavelength has been measured the interferometer can be used to measure arbitrary distances.

Additionally the frequency of the radiation can be measured by means of a cavity wavemeter calibrated by reference to a crystal-controlled oscillator. In the present experiment the accuracy of this calibration was limited by the setting accuracy of the wavemeter to be ± 1 Mc/s. at 24,000 Mc/s. This accuracy could be improved, but it was considered adequate for the present experiments. From a measurement of the frequency made in this way, and the wavelength as determined by the interferometer, the velocity of electromagnetic waves in air can be computed.

In the actual determination of wavelength 80 minima are counted as M1 moves over its traverse of 50 cm. The positions of the minima at the ends of this traverse are determined either by setting on the actual minimum or by setting slightly off the minimum on either side of the response where a finite slope exists. In the latter the setting accuracy is better, but care must be taken that the response is symmetrical or erroneous results may be obtained.

Using transmitted powers of a few milliwatts, and with the component parts of the interferometer spaced at distances not exceeding 2 metres from the beam divider, the measurement of the rectified current through a 12·5 mm. crystal detector in the waveguide feed from the receiving horn has proved adequate, and currents from 200–400 microamps have been obtained. Maximum to minimum ratios of this current up to 10,000 : 1 have been noted in some instances. Due to diffraction effects there is, as mentioned earlier, some variation in the maximum to minimum ratio of this current as the mirror M1 is moved. Nevertheless, using a spot galvanometer of 20 ohms resistance and a full-scale deflection of 8 microamps, setting accuracies on the minimum within 10^{-3} cm. can be achieved in most instances when using the 6 in. × 6 in. transmitting and receiving apertures; by setting on the sides of the response the setting accuracy approaches 10^{-4} cm., the limit being set by slight power output fluctuations of the oscillator value.

For wider spacings of the interferometer the sensitivity using the crystal detector and galvanometer technique is inadequate, and a 12·5 mm. superheterodyne receiver is used. The local oscillator of this receiver is locked to a cavity by means of another Pound Frequency Stabilizer, in order to keep the receiver in tune. In spite of the large gain available in such a receiver it has not been found possible to improve on the setting accuracy obtained with the crystal and galvanometer technique. This is due to the effects of diffraction and stray radiation in the present interferometer.

3.3. *Measurement of Dielectric Constants*

Using the optical analogy it would appear that the dielectric constant of a plane sheet of low-loss material is determined simply by inserting a sheet of the material normally into one of the beams of the interferometer and measuring the resultant shift of a minimum. The ambiguity as regards the number of minima in the actual shift would have to be resolved by using sheets of different thickness, unless the magnitude of the dielectric constant is known approximately. The refractive index, and hence the dielectric constant of the material, is given by

$$t(\mu-1)=x, \qquad \ldots\ldots(3)$$

t being the thickness of the sheet, μ the refractive index, and x the shift of a minimum. This method, even neglecting diffraction effects, can only give an approximate result, since it neglects the reflections which occur at the surfaces and inside the specimen; it will thus give better approximations for low dielectric constants.

Whilst approximate results using this method have been obtained for Perspex and polystyrene sheets, the values of dielectric constant obtained have been found to depend on both the size and the position of the specimen in the beam, this dependence being no doubt connected with the diverging nature of the beams in the interferometer as well as the reflections occurring at the specimen.

To eliminate these uncertainties the sheets were made the same size as the mirrors and were mounted directly on to mirror M1. In this position the method becomes analogous with the short-circuited line technique (von Hippel and Roberts 1946) of determining dielectric constants.

The simple formula of equation (3) gives the first approximation, but even with dielectric constants of the order of two, more accurate methods of impedance transformation must be applied. Briefly, measurements of the shift of a minimum on placing the specimen on mirror M1, and the thickness, enable a determination to be made of the distance of the first minimum in air from the dielectric boundary. Then by applying impedance transformations back from this minimum to the dielectric–air surface, and also from the mirror through the dielectric to this interface and equating the two impedances, we obtain a transcendental equation involving the dielectric constant as an unknown. In practice, a simpler method for dielectric materials of low loss is to use the circle diagram of impedances and to perform this matching of impedances by trial and error, using the approximate value for the dielectric constant, equation (3), as a guide.

§ 4. RESULTS

4.1. *Spacing of the Interferometer*

Measurements have been made at five positions of the interferometer. The respective distances in centimetres of the component parts from the beam divider are shown in Table 1.

Table 1

Position	1	2	3	4	5
Transmitting horn	150	100	50	25	310
Receiving horn	150	100	50	25	150
Mirror M1	175	125	75	50	600
Mirror M2	175	125	75	50	40

The distance of mirror M1 from the beam divider was measured when M1 was at the centre of its traverse, i.e. when set at 25 cm. on the 50 cm. scale.

4.2. *Wavelength Measurements*

Measurements of the apparent wavelength were made at each of the positions given in Table 1, using the 6 in. × 6 in. apertures of the receiving and transmitting horns. These measurements were made both by setting on the actual minima and by setting on the sides of the response. The results are shown in Table 2. Deviations from the mean of the measurements made in each position were about one part in 10^4. During all determinations of the wavelength the operating frequency was constant at $24{,}070 \pm 1$ Mc/s., as indicated by the calibrated wavemeter. The lead screw used to move mirror M1 has been checked at 20° c. and found to have an average error of one part in 10^4 over a length of 40 cm. This error has been allowed for in the results quoted in Table 2.

Table 2. Apparent Wavelength

Position		1	2	3	4	5
Wavelength (cm.)	using minimum	1·2452	1·2455	1·2456	1·2459	1·2451
	using response	1·2452	1·2455	1·2456	1·2459	1·2451

Values of the apparent velocity of electromagnetic waves in air at 20° c. have been deduced from the known frequency and measured wavelength of the radiation, and the results are shown in Table 3.

Table 3. Measured Velocity of Electromagnetic Waves

Position	1	2	3	4	5
Velocity (10^{10} cm/sec.)	2·9972	2·9980	2·9982	2·9989	2·9970

Using the present standard value for the velocity of light *in vacuo*, viz. $2{\cdot}99776 \times 10^{10}$ cm/sec., and assuming a value 1·0006 for the dielectric constant of air under the conditions of the experiment, we obtain for the velocity of electromagnetic waves in air a value of $2{\cdot}9969 \times 10^{10}$ cm/sec. to the approximation required, viz. 1 in 10^4. To this approximation the effects of variations in humidity and pressure on the measurements have been neglected. These would need consideration in a more accurate determination.

From Table 2 it is seen that the value obtained for the wavelength depends on the spacing of the component parts of the interferometer. The measured wavelength increases as the spacing is reduced, this being due no doubt to the spherical wave fronts existing in the interferometer. From Table 3 it is seen that there is agreement, within the limits of experimental error, between the deduced velocity of electromagnetic waves and the generally accepted value, this agreement occurring at the wide spacing of the interferometer in position 5. The operating wavelength as measured in position 5 is 1·2451 ± 0·0001 cm.

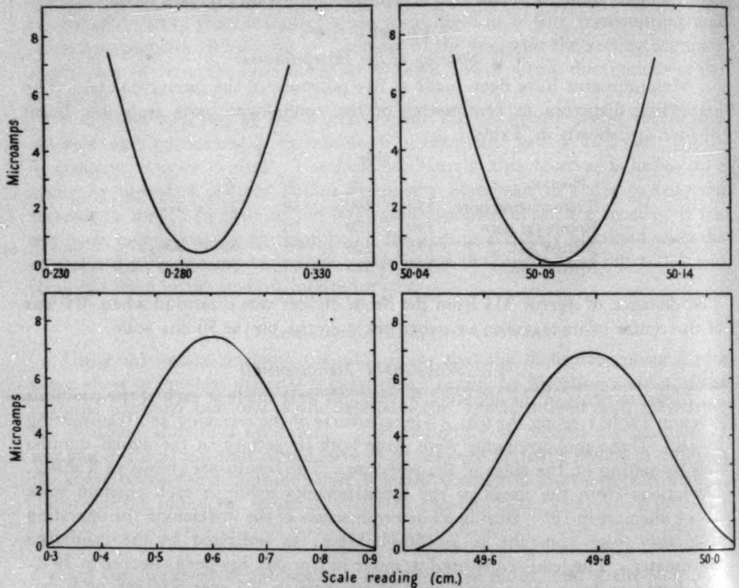


Figure 4. Maximum and minimum responses at position 1; 6 in. × 6 in. apertures.

Measurements of the responses at a maximum and minimum were made with the interferometer in position 1. The results are shown in Figure 4. The symmetry of the minimum response approaches 1 in 10^4, which is about the accuracy of the wavelength determination in position 1.

Figure 5 shows a plot of a number of maxima and minima occurring on the interferometer around a scale reading of 25 cm. for mirror M1, the mirrors then being at equal distances from the beam divider.

Maximum and minimum responses with the interferometer in position 4 are shown in Figure 6. The symmetry of the minimum response no longer approaches 1 in 10^4, and the accuracy of the wavelength measurement at this position is no longer within this limit.

The effect of reducing the 6 in. × 6 in. radiating and receiving apertures was studied in position 4. The apertures were reduced by mounting $\frac{1}{8}$-in. thick brass plates over the ends of the horns, in which apertures of 3 in. × 3 in. and $1\frac{1}{2}$ in. × $1\frac{1}{2}$ in. were made. The wavelength was then measured for the various apertures used, and the effect of these on the wavelength is shown in Table 4. During these experiments it was possible to observe the efficiency of the metal lenses in reducing the beam divergence of the transmitting horn and in increasing the effective receiving area of the receiving horn. Removal of the lens from the transmitting horn reduced the output at a maximum from 100 microamps to 32 microamps, with a similar reduction when the receiving horn lens was removed. Removal of both these lenses caused the maximum to fall from 100 microamps to 16 microamps.

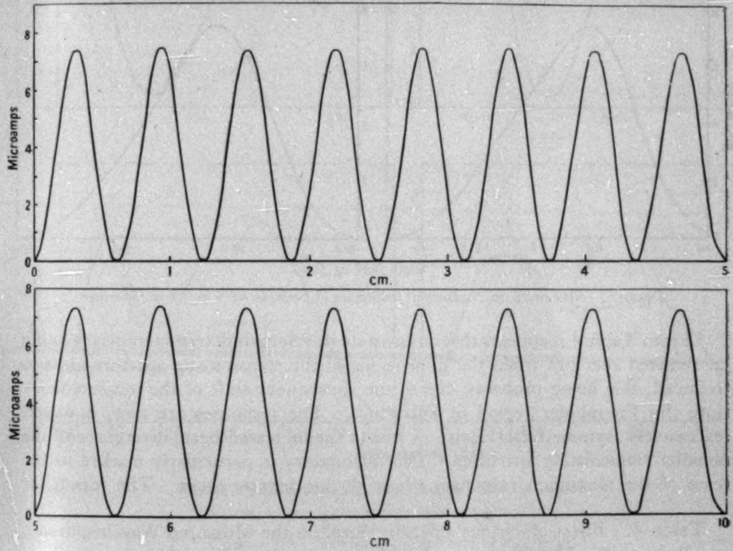


Figure 5. Plot of responses of the interferometer at position 1.

Figure 7 shows the maximum and minimum responses obtained when the transmitting aperture is reduced with the interferometer in position 4; Figure 8 shows the variation in the amplitude of the maxima as M1 moves over its traverse of 50 cm. at positions 1 and 4, and the variation for the various apertures used at position 4. Figure 9 shows the effect of increasing the transmitted power on the sensitivity at the minimum.

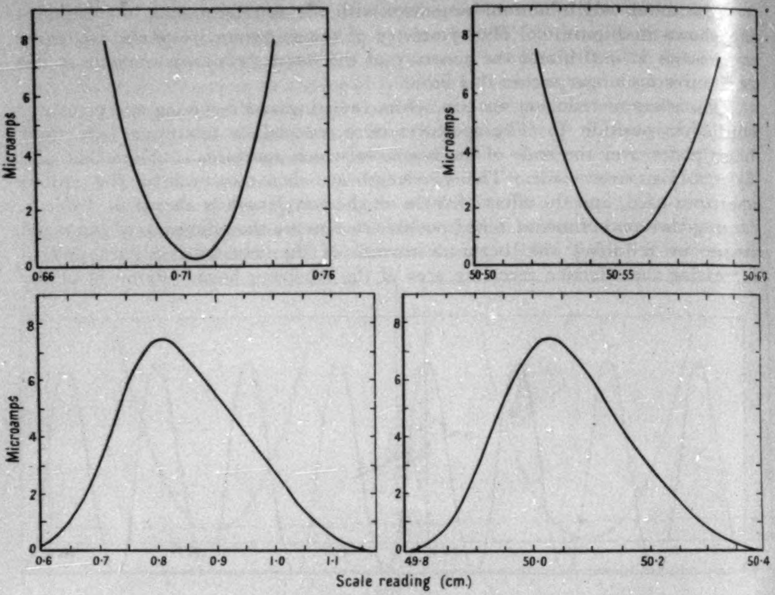


Figure 6. Maximum and minimum responses at position 4; 6 in. × 6 in. apertures.

From Table 4 it appears that an accurate wavelength determination is possible at reduced spacings when the dimensions of the transmitting aperture are also reduced, this being probably due to the consequent shift of the measurements into the Fraunhofer region of diffraction. The responses are now, however, excessively asymmetrical (Figure 7), due to the increased beam divergence of the smaller transmitting apertures. This asymmetry is particularly marked in the case of the maximum responses where double humps occur. The reason for

Table 4. Effect of various Aperture Sizes on the Measured Wavelength at Position 4. Metal lenses used with the aperture unless otherwise stated.

Apertures		Wavelength (cm.)	
Transmitter	Receiver	Using minimum	Using response
6 in. × 6 in.	6 in. × 6 in.	1·2459	1·2459
3 in. × 3 in.	6 in. × 6 in.	1·2452	1·2452
1½ in. × 1½ in.	6 in. × 6 in.	1·2451	1·2448
6 in. × 6 in.	3 in. × 3 in.	1·2460	1·2461
3 in. × 3 in.	3 in. × 3 in.	1·2461	1·2461
6 in. × 6 in.	1½ in. × 1½ in.	1·2457	1·2457
6 in. × 6 in.—lens off	6 in. × 6 in.—lens off	1·2468	
6 in. × 6 in.	6 in. × 6 in.—lens off	1·2454	
6 in. × 6 in.—lens off	6 in. × 6 in.	1·2453	
3 in. × 3 in.—lens off	6 in. × 6 in.	1·2461	
1½ in. × 1½ in.—lens off	6 in. × 6 in.	1·2462	

these is not clear; they do not seem to be due to the secondary lobes in the polar diagram of the transmitting aperture; presumably they are due to reflection effects arising at the conducting plates used to reduce the transmitting aperture.

Use of the interferometer for the measurement of small frequency changes has also been made in position 5. The change in the position of a minimum due to a small frequency change depends on the relative distances of the two mirrors from the beam divider. Actual measured shifts with the interferometer in position 5 amounted to about 1 millimetre for a 10 Mc/s. change in operating frequency; this is in fair agreement with the theoretical value for this position assuming plane wave trains in the interferometer.

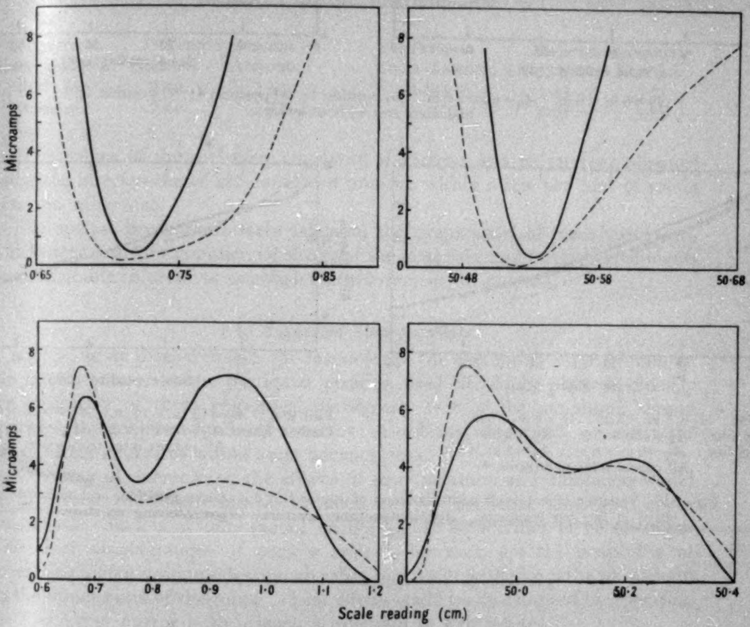
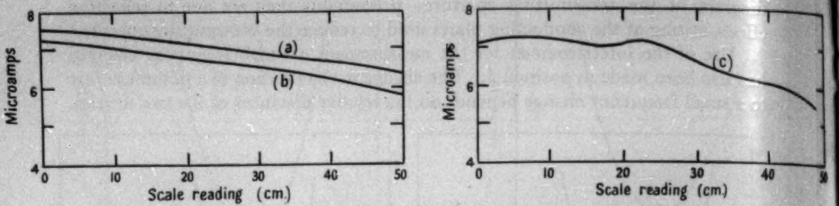


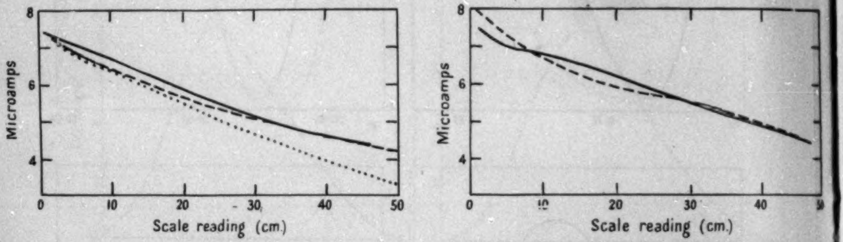
Figure 7. Maximum and minimum responses at position 4.
Curve ——— Transmitting aperture 3 in. × 3 in.
Receiving aperture 6 in. × 6 in.
Curve - - - - Transmitting aperture 1½ in. × 1½ in.
Receiving aperture 6 in. × 6 in.

4.3. *Dielectric Constant Measurements*

Measurements of the dielectric constants of Perspex and polystyrene have been made in order to assess the value of the interferometer for this purpose. Two values of thickness of material were used, and measurements were made at all positions of the interferometer as well as with the various transmitting and receiving apertures used in position 4. The values quoted in Table 5 are those



A_{T}=6 in.×6 in. A_{R}=6 in.×6 in. (*a*) position 1; (*b*) position 4; (*c*) position 4; both apertures without lenses.



Full line: A_{T}=3 in.×3 in. A_{R}=6 in.×6 in.
Dotted line: A_{T}=3 in.×3 in. A_{R}=3 in.×3 in.
Dashed line: A_{T}=6 in.×6 in. A_{R}=3 in.×3 in.
All curves refer to position 4.

Full line: A_{T}=1½ in.×1½ in. A_{R}=6 in.×6 in.
Dashed line: A_{T}=6 in.×6 in. A_{R}=1½ in.×1½ in.
All curves refer to position 4.

Figure 8. Variation of amplitude with transverse of mirror 1 at the various positions and apertures used in the interferometer. A_{T}=transmitting aperture; A_{R}=receiving aperture.

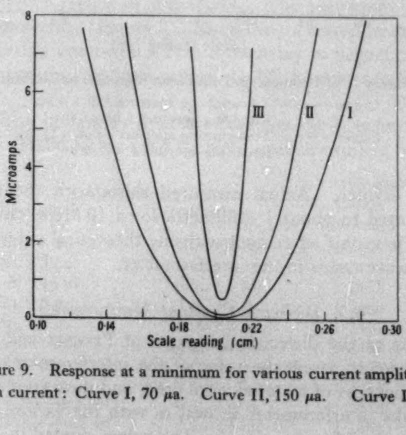


Figure 9. Response at a minimum for various current amplitudes.
Maximum current: Curve I, 70 μa. Curve II, 150 μa. Curve III, 720 μa.

obtained at position 1 using the 6 in. × 6 in. apertures. The maximum deviation from these values over the full range of measurements amounted to about 1%, so that presumably the 'edge effect', due to the sheet being the same size as the mirror, is small.

Table 5. Dielectric Constants of Polystyrene and Perspex at 24,000 Mc/s. measured on the Interferometer

Polystyrene Sheet thickness	Dielectric constant Approx.	Accurate	Perspex Sheet thickness	Dielectric constant Approx.	Accurate
0·212 in.	2·43	2·63	0·183 in.	2·45	2·71
0·382 in.	2·66	2·63	0·252 in.	3·05	2·76

The values of the dielectric constants of Perspex and polystyrene obtained using the interferometer are consistent and are within a few per cent of values obtained otherwise.

No special precautions were taken in the preparation of these specimens, which were ordinary commercial sheets of the materials, and further experiments seem desirable in order to establish the accuracy of the method.

4.4. *Suggested Improvements*

For accurate measurements the responses of the maximum and minimum on the interferometer at the particular position used should be plotted. Ideally the symmetry of these responses, particularly that of the minimum, should approach the required limits of accuracy; in order to obtain this symmetry the use of larger apertures would seem advantageous.

One way of overcoming the effects of stray radiation and diffraction would be to enclose the interferometer in metal tubes of sectional dimensions about 6 in. × 6 in. Such a system should enable higher sensitivities to be obtained. The chief disadvantages of such a system, however, are the possibility of generating higher order modes in such tubes, and the dependence of the wavelength on the dimensions of the tube. These effects could be investigated by operating the measuring mirror both in such a tube and in free space.

There are distinct possibilities of increasing the accuracy of the wavelength measurement, and hence of the determination of the velocity of electromagnetic waves, but it is clear that a more accurate and robust interferometer would be required, together with accurate control of temperature, pressure and humidity.

§ 5. USE AS A SUBSTANDARD OF LENGTH

The application of the interferometer for the accurate calibration of relatively large lengths seems possible. For this purpose adequate discrimination at the minimum, when large path differences exist in the interferometer, would be required.

From the results obtained in position 5, where the path difference was about 12 metres, it seems clear that the conditions for adequate discrimination, namely high-frequency stability of the source and beams of equal intensity, can be met.

In order to measure a length of 5 metres to one part in 10^6 the wavelength, as determined by the interferometer in the particular position used, would have to remain constant over this length to one part in 10^6 (cf. diffraction effects).

To verify this it would be necessary to increase the accuracy of the wavelength determination by a factor of 100. A better procedure would be to calibrate the interferometer in the particular position used against a standard of length, when a setting accuracy of 10^{-4} cm. would be required for lengths of 5 metres.

ACKNOWLEDGMENTS

The author would like gratefully to acknowledge the assistance given by Mr. J. N. Cruickshank in the design and operation of the Pound Frequency Stabilizer. The work described in this paper was carried out for the Department of Scientific and Industrial Research at the Telecommunications Research Establishment, Ministry of Supply. Acknowledgment is made to the Chief Scientist, Ministry of Supply, the Director of Radio Research, Department of Scientific and Industrial Research, and to the Controller, H.M. Stationery Office, for permission to publish the paper.

REFERENCES

von Hippel, A., and Roberts, S., 1946, *J. Appl. Phys.*, **17,** 610.

Kock, W. E., 1946, *Proc. Inst. Radio Engrs.*, **34,** 828.

Pound, R. V., 1946, *Rev. Sci. Instrum.*, **17,** 490, *Proc. Inst. Radio Engrs.*, **35,** 1405.

Stratton, J. A., 1941, *Electromagnetic Theory* (London and New York: McGraw Hill Book Co.), § 8.13.

Wood, R. W., 1934, *Physical Optics* (New York: Macmillan Book Co.), p. 292.

Note on the Focusing of Electron Beams in certain Magnetic Fields

By P. A. STURROCK

National Bureau of Standards, Washington, D.C.

Communicated by L. Marton; MS. received 13*th March* 1950, *and in amended form* 3*rd July* 1950

ABSTRACT. Equations are set out which determine the focusing properties of electron beams in magnetic fields whose scalar potential has a plane of antisymmetry. From these is derived the condition that a proposed ray-axis and associated focusing requirements should be physically realizable. It is also shown that the fringe-effect of fields with sharply defined boundaries may be characterized by a pair of focal lengths for which formulae are given.

§ 1. INTRODUCTION

Magnetic fields whose scalar potential possesses a plane of antisymmetry, the ' equatorial ' plane, have long been used in mass spectrometers and beta-ray spectrometers (Siday 1947). Such fields are often prescribed to be constant over the plane within defined boundaries; ray-tracing in the equatorial plane is then simplified; moreover, such fields may be approximately realized by magnets having flat-faced pole pieces. The fringe effect is usually treated as a correction (Coggeshall 1947).

§ 2. THE PARAXIAL EQUATIONS

Consider a magnetic field whose scalar potential has a plane of antisymmetry; the field is then determined by the (normal) field strength H in the equatorial plane. Let a particular ray in the equatorial plane be the ' ray-axis '. We may then give to a point P coordinates (u, v, w), where $w = PM$, M being the foot of the normal from P to the equatorial plane, $v = MN$, N being the foot of the normal from M to the ray-axis, and u measures the arc-length of N along the ray-axis (Figure 1).

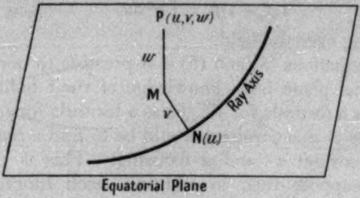


Figure 1.

If $\sigma(u)$ is the curvature of the ray-axis, measured so as to be positive when the centre of curvature lies on the side of the ray-axis for which v is positive, then

$$\sigma = H^0/p, \qquad \ldots\ldots(1)$$

where $H^0(u)$ is the field strength on the ray-axis, measured in the direction of the w-axis, and p is the $H\rho$ momentum of the beam.

Rays within a certain neighbourhood of the ray-axis may be characterized by equations between their coordinates of the form $v = v(u)$ and $w = w(u)$. Then for rays indefinitely close to the ray-axis, the functions $v(u)$ and $w(u)$ are solutions of the following second-order differential equations:

$$\frac{d^2v}{du^2} + Vv = 0, \quad \text{wherein } V = (H^0/p)^2 - H^0{}_v/p, \qquad \ldots\ldots(2)$$

and

$$\frac{d^2w}{du^2} + Ww = 0, \quad \text{wherein } W = H^0{}_v/p, \qquad \ldots\ldots(3)$$

the function $H^0{}_v(u)$ being the gradient of H in the v direction, evaluated on the ray-axis.

The equations (2) and (3) are formally identical with the equations determining the paraxial behaviour of rays in an optical system with two orthogonal planes of symmetry. It is seen at once that if $H^0{}_v$ is zero, as in a uniform field, then V is positive and W is zero, so that there is convergent focusing of rays in the equatorial plane but no focusing of rays diverging from the plane.

If the field is confined to a small region along the ray-axis, we may deduce from equations (2) and (3) that the focusing action of the field may be represented by a pair of focal-lengths, f_v and f_w, where

$$1/f_v = \int V du \quad \text{and} \quad 1/f_w = \int W du, \qquad \ldots\ldots(4)$$

the integration extending over the field.

§ 3. CONDITION FOR A PROPOSED BEAM TO BE PHYSICALLY REALIZABLE

The necessary and sufficient condition that a ray-axis, characterized by a function $\sigma(u)$, and an associated assembly of the neighbouring rays, characterized by functions $V(u)$ and $W(u)$, should be physically realizable is that

$$V + W = \sigma^2. \qquad \ldots\ldots(5)$$

Examination of equations (1), (2) and (3) shows that this is necessary; it is also sufficient, for, if the condition is satisfied, values of $H^0(u)$ and $H^0{}_v(u)$ may be determined which will realize the rays.

It follows at once that, for a short field,

$$1/f_v + 1/f_w = \int \sigma^2 \, du, \qquad \ldots\ldots(6)$$

the integral extending over the field.

From either of equations (5) and (6) it is possible to derive the behaviour of rays *off* the equatorial plane from knowledge of their behaviour *in* the plane: (5) may be treated as a formula for W, (6) as a formula for f_w.

A simple example of this corollary would be to find a realizable ray-assembly which achieves concurrent v- and w-focusing. This is certainly obtained if $V = W$ for all u: suppose that, in addition, each function is constant. If $V = W = 1/2a^2$, equation (5) asserts that the rays will be realizable if, and only if, $\sigma = 1/a$. We see that the ray-axis is a circle of radius a and that any two points on the circumference subtending an angle $\pi\sqrt{2}$ to the centre are ‘doubly’ conjugate. Equations (1) and (3) show that H^0 and $H^0{}_v$ should have the values p/a and $p/2a^2$, respectively. The system is recognized to be the ‘double-focusing’ beta-ray spectrometer of Svartholm and Siegbahn (1947).

§ 4. THE FRINGE-EFFECT

Suppose that H has the value H^* over a certain domain of the equatorial plane, outside of which it falls rapidly to zero. The ‘fringe-effect’ is the effect of the field outside the domain boundary (which, in practice, will approximate to the contours of the pole pieces). It is assumed that the extent of the fringe field is short compared with the radius of curvature of the ray-axis.

Suppose that the ray-axis intersects the field-boundary at O (Figure 2), making an angle θ with the normal at O directed inwards. If H_{n} is the field derivative along the normal, we see that, in the neighbourhood of O,

$$H^0{}_v = H_{\mathrm{n}} \sin\theta. \qquad \ldots\ldots(7)$$

Moreover, an integral along the ray-axis may be transformed into an integral along the normal by the relation

$$dn = du \cos\theta. \qquad \ldots\ldots(8)$$

It is clear from equations (1) and (8) that, if ψ represents the deflection of the ray-axis by the fringe-field,

$$\psi = \sec\theta \int_{\mathrm{fr}} (H/p) \, dn. \qquad \ldots\ldots(9)$$

The earlier assumption about the extent of the fringe-field implies that ψ is small.

The focusing action of the fringe-field may be evaluated by equations (4) together with equations (7) and (8). We find that the effect of the fringe field may be characterized by two focal-lengths, given by the formulae

$$1/f_v = \sec\theta \int_{\mathrm{fr}} (H/p)^2 \, dn - \tan\theta (H^*/p), \qquad \ldots\ldots(10)$$

and

$$1/f_w = \tan\theta (H^*/p). \qquad \ldots\ldots(11)$$

Of the two terms making up formula (10), the former represents the effect of the finite extension of the fringe-field, the latter the effect of the inclination of the beam to the field-boundary; only the latter effect appears in formula (11). The latter term of formula (10) is taken into account in ray-tracing, so that it is the former term only which is usually interpreted as 'fringe-effect'. Since, for a narrow fringe, this term is small, the most important fringe-effect is seen to be

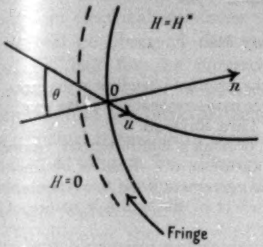


Figure 2.

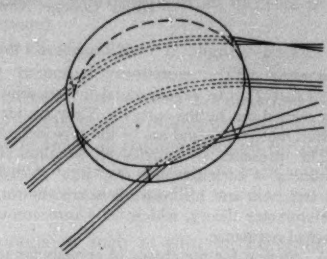


Figure 3. Focusing in w-direction due to the fringe-field of circular pole pieces.

that represented by formula (11). Figure 3 demonstrates how the fringe-focusing in the w-direction depends on the angle of intersection of the beam and the field-boundary.

§ 5. CONCLUSION

It is hoped that this note will prove useful in visualizing the focusing of the kind of fields discussed, and in picturing the behaviour of beams in fringe-fields. Neglect of the w-focusing of the fringe-field of a beta-ray spectrometer will, in general, result in heavy loss of beam current; by taking this focusing into account, transmission would be substantially improved.

These results are extracted from an analysis of the image-forming and image-marring properties of arbitrary magnetic fields, which it is hoped to publish at a later date. It has been pointed out to the author that the substance of the note is also implicit in the work of Cotte (1938).

REFERENCES

Coggeshall, N. D., 1947, *J. Appl. Phys.*, **18**, 855.
Cotte, M., 1938, *Ann. Phys., Paris*, **10**, 333.
Siday, R. E., 1947, *Proc. Phys. Soc.*, **59**, 905.
Svartholm, N., and Siegbahn, K., 1947, *Ark. Mat. Astr. Fys.*, **33A**, No. 21.

Satellite Resonances in Ultrasonic Interferometry

By J. F. W. BELL
Kings' College, Newcastle-upon-Tyne

MS. received 14*th February* 1950

ABSTRACT. An experimental investigation into the origin of satellite resonances has been carried out by the author. They are identified as mode resonances of the gas in the interferometer tube and are of the type described by Lord Rayleigh.

The presence of unresolved satellites in the principal interferometer resonance introduces a considerable error into absorption measurements. Results obtained by Van Itterbeek and his co-workers are shown to be in agreement with the Krasnooshkin interferometer theory, which takes into account the effect of the multiple nature of the principal resonance.

A criterion for the choice of crystals for ultrasonic absorption measurements is given.

PART I.—THE ORIGIN OF SATELLITE RESONANCES

§ 1. INTRODUCTION

An effect frequently encountered in the use of the ultrasonic interferometer for measurements on gases is the occurrence of more than one series of resonances. The extra resonances fall into two categories, namely those which are due to lack of parallelism between crystal and reflector, and the rest which are characteristic of the crystal and interferometer tube used. The former are eliminated when crystal and reflector are exactly parallel.

The latter, which have been widely discussed in the literature, have been the subject of the experimental investigation described in this paper. The original description of these resonances, which, because of their proximity to the principal resonance, are known as satellites, was given by Pielemeier (1931). He deduced that the presence of more than one series of resonances indicated that the ultrasonic energy was being propagated with more than one velocity. This he classified with the variation of velocity with intensity observed near explosions.

It was soon evident that this explanation could not be correct as the character of the resonances did not change with the power supplied to the generator.

In a second paper Pielemeier (1938) pointed out this error in his previous work and suggested that more than one frequency was present in the vibrating crystal. These other frequencies gave rise to differently spaced resonances. It is very evident that such an explanation is inapplicable to the case where the crystal is made to resonate by impressing a voltage. In the steady state the crystal will vibrate at the applied frequency only.

§ 2. EXPERIMENTAL INVESTIGATION

Preliminary measurements on air were carried out with a $1\frac{1}{2}$-in. diameter cylindrical interferometer using a 1-in. diameter X-cut quartz crystal operating at a frequency of 250 kc/s. A radio-frequency bridge was developed to measure the resistive component of the equivalent electrical circuit of the crystal. A part of this resistance arises from the acoustic radiation from the vibrating faces of the crystal. Figure 1 shows the variation of electrical resistance of the crystal with the length of the column of air between the crystal and reflector.

At resonances, which occur every half-wavelength, the coupling between the crystal and the column of air increases. This results in an increase in the equivalent electrical resistance of the crystal.

It will be observed that the length of air column is slightly greater for satellites than for the corresponding principal resonance. The separation increases as the order of resonance increases. On plotting the distances of the satellites from the principal against the length of the column corresponding to the principal resonance, a series of straight lines, all passing through the origin, is obtained (Figure 2). Thus a wavelength λ_s can be ascribed to each satellite. The slopes of the lines give the fractional excess of satellite wavelength over principal wavelength λ.

To discover the source of a phenomenon such as this, an experimental approach often used is to vary the conditions and instrument parameters as far as possible. If circumstances can be found which modify the phenomenon in any way, a valuable clue to its origin is obtained; at the least this will give rise to the deduction of a positive experimental law of behaviour.

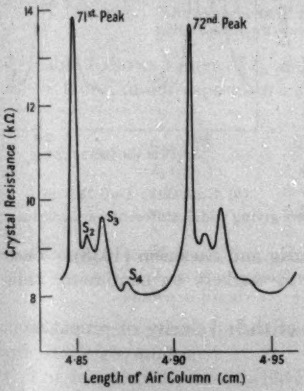


Figure 1. Satellite patterns for air.

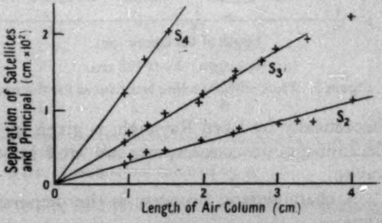


Figure 2. Plots of the distances of the satellites from the principal resonance as the length of the column of air is varied.

The first change attempted was to line the walls of the interferometer tube with a layer of cotton wool in order to damp out any reflection from the side walls. The results were inconclusive in that all resonances were reduced in strength, but no change in wavelength could be detected.

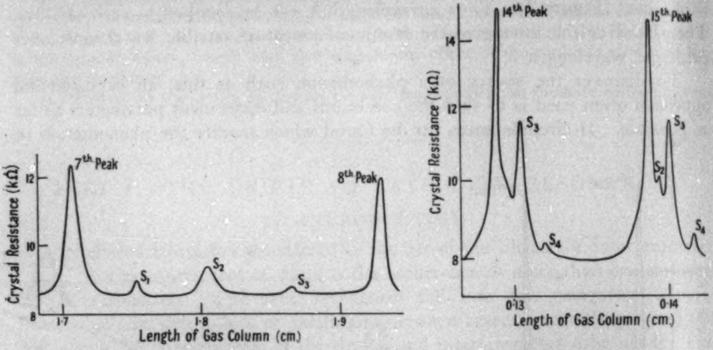
It was next decided to investigate the effect of change of principal wavelength. Selection of gases giving a wide range of velocities of sound, together with low absorption (to give good resolving power) was necessary. Hydrogen has by far the greatest velocity of sound of the gases readily obtainable; in addition, its absorption coefficient is small. For velocities intermediate between air and hydrogen, mixtures of air, hydrogen and coal gas were convenient. For velocities less than that of air no gas considered was suitable as, although the velocity in polyatomic gases is low, the absorption, because of molecular effects, is high.

As it was found that the satellites were very close to the principal at the shorter wavelengths, no observations on low velocity gases were attempted.

Figure 3 shows a series of observations for different gases and gas mixtures.

§ 3. THEORETICAL CONSIDERATIONS

In a recent paper by Krasnooshkin (1944) on the theory of the interferometer it was pointed out that the non-uniform vibration of the crystal surface could result in the excitation of transverse modes of resonance in an interferometer tube. An excellent account of these modes at audio frequencies, first discovered theoretically by Lord Rayleigh, is given by Hartig and Swanson (1938). Their electromagnetic counterparts are well known to workers on centimetre radio waves.



(*a*) Hydrogen: $\lambda=0{\cdot}478$ cm. (*b*) Coal Gas: $\lambda=0{\cdot}183$ cm.

Figure 3. Plots of the satellite resonances for the two gases giving widely different wavelengths.

A characteristic property is the dispersion of their velocity of propagation. Thus

$$\lambda_s=\lambda\Big/\left\{1-\left(\frac{\lambda}{\lambda_0}\right)^2\right\}^{1/2}. \qquad \ldots\ldots(1)$$

Here λ is the wavelength for plane waves (the principal wavelength is assumed to be equal to λ) and λ_0 is a function of the tube diameter d, which was equal to 3·8 cm., and the mode:

$$\lambda_0=\pi d/x_{MN}. \qquad \ldots\ldots(2)$$

x_{MN} is the Mth root of the first derivative of the Bessel function of the first kind of order N. For complex modes x_{MN} is large and λ_0 is small. When $\lambda=\lambda_0$, λ_s is infinite; at this wavelength the phase velocity is infinite and energy can no longer be propagated in the mode; λ_0 is known as the cut-off wavelength for the particular mode.

In the experimental case investigated $\lambda \ll \lambda_0$. Thus equation (1) approximates to

$$\frac{\lambda_s-\lambda}{\lambda}=\frac{1}{2\lambda_0{}^2}\lambda^2. \qquad \ldots\ldots(3)$$

If satellites are mode resonances their wavelengths must follow this law.

§ 4. INTERPRETATION OF RESULTS

Figure 4 shows $(\lambda_s-\lambda)/\lambda$ plotted against λ^2. The straight lines obtained show that equation (3) is followed quite well. It only remains to see whether reasonable values of x_{MN} are obtained from these observations.

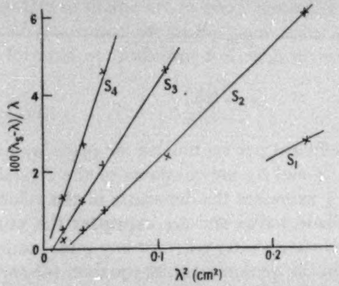


Figure 4. The experimental test of equation (3). The slopes of these lines give the cut-off wavelengths of the modes of vibration.

Table 1 shows values of λ_0 and $\pi d/\lambda_0$ obtained. In every case a value of x_{MN} can be found to correspond to $\pi d/\lambda_0$; the values of M and N are also shown.

Table 1

Satellite	S_1	S_2	S_3	S_4
λ_0 (cm.)	1·85	1·32	1·04	0·91
$\pi d/\lambda_0$	6·4	9·0	11·4	13
x_{MN}	6·7	8·5	11·7	13·3
N	2	1	1	0
M	2	3	4	4

λ_0 is the cut-off wavelength for the various satellites shown in Figure 3. $\pi d/\lambda_0$ is the experimental value of x_{MN}.

There thus remains little doubt as to the identification of satellite resonances with Rayleigh modes.

The lowest order modes have not been detected, presumably because they are so close to the principal resonance that they are not resolvable. The measurements on hydrogen confirm this opinion, for a satellite is observed which was unresolvable in the case of air.

The compounding of the principal resonance with low order modes gives a possible explanation of the fact that the lines of Figure 4 do not pass quite through the origin. The wavelength of the principal resonance has been taken as λ as the wave front was assumed to be plane. If it is actually composed of a number of modes, each having a wavelength greater than λ, the observed wavelength will be slightly in excess of λ. Thus the observed values of $(\lambda_s-\lambda)/\lambda$ will be less than the true values. This would result in a general raising of the lines of Figure 4.

PART II.—THE INFLUENCE OF UNRESOLVED SATELLITES ON ABSORPTION MEASUREMENTS

§ 5. THEORY

Krasnooshkin (1944) has taken into account the fact that the principal resonance peak may consist of a number of low order modes rather than the simple plane waves assumed hitherto. He showed that under these conditions the observed absorption α_{obs} is made up of two factors : (i) the true absorption of acoustic energy by the modes making up the principal resonance—this is slightly greater than the value for plane waves ; (ii) the interferential absorption α_{I} which is due to interference effects between the various modes: this effect is expected to be quite large.

$$\alpha = \frac{\alpha_{\text{obs}} - \alpha_{\text{I}}}{1 + \chi\lambda^2/4\pi^2} \qquad \ldots\ldots(4)$$

$$\alpha_{\text{I}} = \frac{\Delta\chi\lambda}{16\pi} \qquad \ldots\ldots(5)$$

α is the absorption coefficient per centimetre for plane waves in the gas, λ is the free space wavelength, χ and $\Delta\chi$ are constants of the crystal and interferometer of dimensions cm^{-2}. χ expresses the departure of the velocity of the combined modes from that of plane waves and $\Delta\chi$ expresses the multiple nature of the resonance; for a plane wave $\chi = \Delta\chi = 0$; for any pure mode $\Delta\chi = 0$.

The $\chi\lambda^2/4\pi^2$ term in the denominator of equation (4) corrects the absorption of the complex mode system to that of plane waves. As it amounts to only a few per cent it will be neglected.

§ 6. EXPERIMENTAL TESTS

It has recently been pointed out by Kittel (1946–7) that the measurement of low absorption coefficients—those of monatomic and diatomic gases—is very unsatisfactory. There is a lack of correlation between values obtained at different frequencies and by different observers. In addition the values are in excess of the formula of Stokes and Kirchhoff giving the absorption due to viscosity and thermal conductivity.

The Krasnooshkin theory at once suggests the reason for this: what is actually being measured is the sum of α and α_{I}; as the value of α_{I} will be different for each crystal used the divergence obtained is not surprising.

A method of separating the two absorptions is available as α is inversely proportional to the pressure. Expressing equation (4) in the form necessary for the determination of α_{I}

$$\alpha_{\text{obs}} = \frac{\alpha_0}{p/p_0} + \alpha_{\text{I}} \qquad \ldots\ldots(6)$$

where p is the pressure and α_0 the gas absorption corresponding to atmospheric pressure p_0. A plot of α_{obs} against $1/p$ will give a straight line of slope $\alpha_0 p_0$ and intercept α_{I} (see Figure 5). If, using the same crystal and interferometer, a determination of α_{I} is made for gases of different velocities equation (2) can be tested.

Measurements suitable for this test have been carried out by Van Itterbeek and his co-workers (Van Itterbeek and Mariens 1937, Van Itterbeek and Thys 1938) using a 304·4-kc/s. crystal. The use of oxygen and hydrogen at a number of temperatures gave a wide range of velocities.

Table 2 summarizes the results deduced from these data, and graphs of (α_{obs}, $1/p$) are shown in Figure 5.

The accuracies of values of α_0 given in the last column of Table 2 are calculated on the assumption that the measured values α_{obs} were accurate to ± 5%.

Plotting α_I against λ (Figure 6) gives a straight line passing through the origin, showing that equation (5) is followed. The value of $\Delta\chi$ deduced was 35 cm^{-2}.

§ 7. CONCLUSION

The experimental data available are in good agreement with Krasnooshkin's theory. The values of absorption are of the same order of magnitude as those calculated from the formula of Stokes and Kirchhoff, but the accuracy leaves much to be desired because of the masking effect of a large value of α_I. Crystals with a simple vibration of the radiating surface, the ideal being piston-like motion, give the low values of α_I desirable. A criterion for the selection of crystals as ultrasonic generations is thus available. Considerable work on the measurement of $\Delta\chi$ for crystals of different sizes and frequencies, leading to an accurate determination of the absorption coefficients of gases of the type mentioned above is still to be carried out.

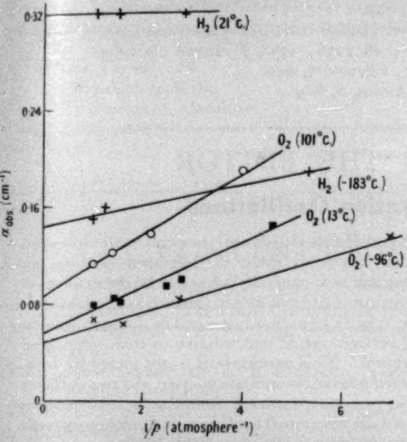


Figure 5. The graphs show the variation of measured absorption coefficient with pressure for gases of various velocities.

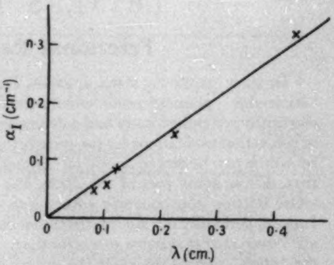


Figure 6. α_I is the absorption coefficient that would be observed if a gas of zero absorption coefficient and wavelength λ were used in the interferometer.

Table 2. Interpretation of the Results of Van Itterbeek's Measurements in Terms of Krasnooshkin's Theory

Gas	Temperature (° c.)	λ (cm.)	α_I (cm⁻¹)	α_0 (cm⁻¹)	α_0 (calc.) (cm⁻¹)	Estimated error in α_0 (cm⁻¹)
Oxygen	110	0·122	0·084	0·026	0·032	±0·005
Oxygen	13	0·107	0·057	0·017	0·032	±0·004
Oxygen	− 96	0·083	0·049	0·013	0·025	±0·003
Hydrogen	12	0·436	0·32	0·000	0·015	±0·015
Hydrogen	−183	0·226	0·145	0·008	—	±0·007

α_0 (calc.) gives the values calculated from the formula of Stokes and Kirchhoff.

ACKNOWLEDGMENTS

The work was carried out at University College, Hull. The author wishes to thank Professor L. S. Palmer for placing the research facilities of the Physics Department at his disposal, Mr. J. M. Hough for his discussions and assistance and Mr. Taylor for construction of the interferometer.

REFERENCES

Hartig, H. E., and Swanson, A. W., 1938, *Phys. Rev.*, **54,** 618.
Kittel, C., 1946–7, *Rep. Prog. Phys.*, **11,** 216 (London : Physical Society).
Krasnooshkin, P. E., 1944, *Phys. Rev.*, **65,** 190.
Rayleigh, Lord, 1900, *Theory of Sound*, Vol. II (London : Macmillan and Co.), p. 279.
Pielemeier, W. H., 1931, *Phys. Rev.*, **37,** 1682, **38,** 1236 ; 1938, *J. Acoust. Soc. Amer.*, **9,** 212.
Van Itterbeek, A., and Mariens, P., 1937, *Physica*, **4,** 609.
Van Itterbeek, A., and Thys, L., 1938, *Physica*, **5,** 889.

LETTERS TO THE EDITOR

Frictional Relaxation Oscillations

In their paper on static friction, Parker and Hatch (1950) seek to explain the so-called ‘ stick–slip ’ phenomenon—encountered when frictional force is measured between an elastically restrained body and a driven surface, and now generally agreed to be the exhibition of relaxation oscillations by the measuring instrument and not due to the intrinsic mechanism by which two bodies in frictional contact slide relative to each other—and in doing so assume that, during some part of the cycle, the two surfaces are at rest relative to each other, and static friction consequently comes into operation. This assumption is not generally true. Indeed, generally during the frictional generated relaxation oscillation cycle, the two surfaces are never at rest relative to each other. This is readily apparent from the fact that during the ‘ stick ’ part of the cycle the dragged elastically restrained surface never catches up with the dragging surface, always having a velocity less than it (Bristow 1942, 1947).

This erroneous assumption is all too commonly made, and has been engendered by the use of the apparently realistic, but actually unrealistic and misleading, term ‘ stick–slip ’.

Dudley and Swift (1949) have recently given an elegant analysis of the motion of elastic restrained body subject to frictional traction by contact with a second body in a large variety of friction–velocity relationships and, attempting to explain the writer's observations, only deduce the exhibition of relaxation oscillations if a period of rest is encountered during which static friction becomes operative.

Neither of these two papers therefore gives satisfactory explanation of frictional relaxation oscillation and consequently gives no clue as to the nature and mechanism of either ‘ dry ’ or ‘ boundary ’ kinetic friction.

The Motor Industry Research Association,
Great West Road,
Brentford, Middlesex.
J. R. Bristow.
19th July 1950.

Bristow, J. R., 1942, *Nature, Lond.*, **149,** 169 ; 1947, *Proc. Roy. Soc.* A, **189,** 88.
Dudley, B. R., and Swift, H. W., 1949, *Phil. Mag.*, **40,** 849.
Parker, R. C., and Hatch, D., 1950, *Proc. Phys. Soc.* B, **63,** 185.

It is clear from Dr. Bristow's letter that our explanation of the so-called ‘ stick–slip ’ phenomenon is not sufficiently full. The behaviour of a soft metal hemisphere in contact with a hardened flat non-metal surface, under the influence of increasing frictional force, will therefore be stated in more detail.

Figure 10 of our paper shows that relative motion takes place between the two surfaces at extremely small tangential loads, and that the total amount of movement is dependent upon the magnitude of the tangential load. This movement is accompanied by an increase in the real area of contact, and when this area cannot grow further, any additional increase in the tangential force gives rise to slip in the ' macro ' sense.

If, now, we imagine these two members to form a ' stick–slip ' pair, that is the soft metal hemisphere to be elastically constrained and the glass member to move at a low uniform velocity, it will be seen that, at the moment of movement, the two surfaces will appear to move together for only a very short time, after which the relative velocity will gradually increase up to the point of macroslip. If it is now assumed that, during the period of macroslip, the two surfaces will become disturbed and present partly or completely new contact areas; this will, in turn, mean that the process will be repeated indefinitely. In this interpretation, it is therefore not necessary to assume that the two bodies are relatively at rest while the tangential force is increasing, that is while movement of the flat is taking place.

Messrs. Ferodo Ltd.,
Chapel-en-le-Frith,
Stockport, Cheshire.
23rd August 1950.

R. C. Parker.
D. Hatch.

CONTENTS FOR SECTION A

PAGE
Prof. P. B. Moon. The Hard Components of Scattered Gamma-Rays . . . 1189
Mr. A. Storruste. The Rayleigh Scattering of 0·41 mev. Gamma-Rays at Various Angles 1197
Prof. S. Devons and Mr. G. R. Lindsey. γ-Radiation from the Resonant Capture of Protons by ^{7}Li Nuclei 1202
Dr. W. J. Swiatecki. The Density Distribution inside Nuclei and Nuclear Shell Structure 1208
Mr. J. Hughes and Dr. K. J. Le Couteur. Spin Orbit Coupling in the Nuclear Shell Model 1219
Dr. K. J. Le Couteur and Mr. S. Zienau. Coherent Scattering of Light by an Atom and Negative Energy States 1223
Dr. J. R. Greening. The Determination of X-Ray Wavelength Distributions from Absorption Data 1227
Dr. S. L. Altmann. Relation between the Franck–Condon Frequencies of Absorption and Fluorescence for some Unsaturated Hydrocarbons 1234
Prof. C. B. A. McCusker. Penetrating Particles in Air Showers 1240
Dr. E. P. George and Mr. J. Evans. Observations of Cosmic-Ray Events in Nuclear Emulsions Exposed below Ground 1248
Prof. L. F. Bates and Dr. J. H. Davis. Heat Changes Accompanying Magnetization in Low and Moderate Fields : the Effects of Strain, and a Theoretical Interpretation 1265
Mr. C. Bull and Dr. G. F. J. Garlick. The Luminescence of Diamonds . . 1283

Letters to the Editor :

Dr. W. E. Moffitt. The Ultra-Violet Spectrum of Ethylene 1292
Mr. R. A. Durie. The Spectra of Flames Supported by Fluorine 1292
Dr. J. B. Birks. Scintillation Efficiency of Anthracene Crystals 1294
Mr. E. E. Salpeter. Dissociation Cross Sections for Fast Hydrogen Molecule Ions 1295
Dr. E. W. Titterton. The Reaction ^{7}Li γp ^{6}He 1297
Mr. P. J. Grant. Forbidden β-Decay in ^{24}Na 1298
Reviews of Books 1300
Contents for Section B 1301
Abstracts for Section B 1301

ABSTRACTS FOR SECTION A

The Hard Components of Scattered Gamma-Rays, by P. B. MOON.

ABSTRACT. A summary is given of theoretical information on the three processes by which gamma-rays can be elastically scattered by atoms, viz. Rayleigh scattering by bound electrons, Thomson scattering by the nuclear charge and (exceptionally) nuclear resonant scattering.

Interference between the three scattered waves is considered and shown to be of practical importance as between Rayleigh and Thomson scattering of hard gamma-rays at large angles.

The calculated intensities of Rayleigh, Thomson and Compton scattering are plotted against angle of scattering for gamma-ray energies of 2·8 and 0·41 MeV., and for scatterers of Al, Cu and Pb. The graphs illustrate the dominance of Rayleigh scattering at very small angles for all energies, and the comparable intensities of Rayleigh and Thomson scattering at high energies and large angles.

The unidentified hard component found by Pollard and Alburger in the large-angle scattering of 2·8-MeV. gamma-rays by various elements is interpreted as a mixture of Rayleigh and Thomson scattering; theoretical and experimental intensities show reasonable agreement for both light (Al) and heavy (Pb) scatterers.

Experiments on the scattering of 0·41-MeV. gamma-rays at about 115° are reported. In agreement with theory, about 2% of the radiation scattered from lead is found to retain the full energy; for copper and aluminium, the proportion of elastically scattered photons is much smaller.

The Rayleigh Scattering of 0·41 *MeV. Gamma-Rays at Various Angles,* by A. STORRUSTE.

ABSTRACT. A report is given on some measurements of the Rayleigh cross section for scattering of 0·41 MeV. gamma-rays by lead, copper and aluminium. At small angles the Rayleigh cross section in lead is found to decrease rapidly with increasing angle of scattering from approximately 12 times the Compton cross section at 3° to a value equal to the Compton cross section at 12°. In copper the Rayleigh cross section is found to equal the Compton cross section at 4·5°. In aluminium no substantial excess scattering above the Compton cross section is found, the smallest angle at which measurements were taken being 3°. At 60°, 90°, 120° and 150° the hard component of the scattered radiation from lead is found to be respectively 1·77, 1·14, 0·88 and 0·66% of the Compton component as calculated by the Klein–Nishina formula.

γ-Radiation from the Resonant Capture of Protons by ^{7}Li Nuclei, by S. DEVONS and G. R. LINDSEY.

ABSTRACT. Further measurements have been made of the angular distribution of the γ-radiation in the reaction $^7\mathrm{Li}(p,\gamma)^8\mathrm{Be}$. The results indicate approximate isotropy, at resonance, for both spectral components (17·6 and 14·8 MeV.) of the radiations. The difficulties in reconciling the properties of the γ-radiation with results of recent experiments on the scattering of protons by lithium are discussed.

The Density Distribution inside Nuclei and Nuclear Shell Structure, by W. J. SWIATECKI.

ABSTRACT. A method of evaluating the density distribution of particles inside a heavy nucleus is described. The results are used to show that the electrostatic repulsion between protons is too weak to produce a nucleus sufficiently hollow to explain the observed ' closed shells ' at N, $Z=50$, 82.

Spin Orbit Coupling in the Nuclear Shell Model, by J. Hughes and K. J. Le Couteur.

ABSTRACT. It is shown that the magnitude of the doublet splitting postulated in Mayer's nuclear shell model is consistent with that observed in ^{5}He. The strength of the indicated spin orbital interaction between pairs of nucleons is worked out.

Coherent Scattering of Light by an Atom and Negative Energy States, by K. J. Le Couteur and S. Zienau.

ABSTRACT. It is proved that the difference between the single electron theory and the positron theory for the coherent scattering of light by an atom represents the scattering of light by the atomic potential. In the practical cases the effect is negligible. The gauge invariance of the perturbation formulae for this effect is discussed.

The Determination of X-ray Wavelength Distributions from Absorption Data, by J. R. Greening.

ABSTRACT. An existing empirical method which may be used, in limited cases, to determine x-ray wavelength distributions from absorption data is given some theoretical foundation. It is shown that the absorption curve function and the wavelength distribution function have the same relationship as a Laplace pair, the absorption curve function being the Laplace transform of the wavelength distribution function. The theory leads to the suggestion of Laplace pairs other than the one previously suggested, thus widening the scope of the method. By using the additive property of Laplace transforms the absorption method of determining spectral distributions is made of general application. The experimental requirements are discussed and examples of wavelength distributions determined by the absorption method are given.

Relation between the Franck-Condon Frequencies of Absorption and Fluorescence for some Unsaturated Hydrocarbons, by S. L. Altmann.

ABSTRACT. Differences between the Franck–Condon frequencies for emission and absorption are calculated for butadiene, naphthalene and anthracene. The molecular-orbital method is used and its validity in the present case is discussed. A correlation with the experimental data has been attempted for anthracene and appears to be satisfactory. It is shown that the calculated values may be correlated qualitatively with the fluorescence efficiencies of the molecules.

Penetrating Particles in Air Showers, by C. B. A. McCusker.

ABSTRACT. The penetrating particles in extensive air showers are compared directly with the penetrating particles in the main cosmic-ray beam. It is found that not more than $70\pm3\%$ of the air-shower particles are single μ-mesons, whilst the remainder are strongly interacting particles. The barometer coefficient of extensive penetrating showers is determined.

Observations of Cosmic-Ray Events in Nuclear Emulsions Exposed below Ground, by E. P. George and J. Evans.

ABSTRACT. Ilford Nuclear research plates were manufactured in a laboratory at an equivalent depth of 60 m. of water below ground. After having been stored at various depths, the plates were processed below ground. The observed frequencies of μ-mesons stopped in the plates are consistent with those expected from measurements on the energy-spectrum of cosmic rays. Several π-mesons stopping in the plates were observed, and reasons for believing these to have been locally produced in the matter near the plates are discussed.

Forty-two nuclear disintegrations have been observed, the frequency at a depth of 60 m. water equivalent being of the order of 5×10^{-3} stars/cm^2/day. Approximately one third of these are attributed to the electromagnetic interaction of μ-mesons on their passage through nuclei. The remainder are attributed in part to neutrons from this first group of stars, and in part to the photons of the soft component underground.

Four examples of stars accompanied by showers of particles at minimum ionization have been observed, and are discussed.

Heat Changes Accompanying Magnetization in Low and Moderate Fields: the Effects of Strain, and a Theoretical Interpretation, by L. F. BATES and J. H. DAVIS.

ABSTRACT. The technique of Bates and his co-workers was used to measure the adiabatic temperature changes occurring during magnetization of nickel under conditions of progressively increased internal strain as a pure specimen was taken from the fully annealed state to the state of maximum strain, in absence of deforming force. A new technique enabled accurate thermal curves during virgin magnetization, and in cycles up to 400 oersteds, to be obtained; the induction effect previously reported was investigated and its origin determined. A new theoretical treatment is described in which account is taken of the energy changes associated with magnetostriction and which enables the thermal curves to be calculated from available data on magnetization and magnetostriction, and a general physical interpretation of the curves is given. Agreement is good for strained specimens and an explanation of the less satisfactory agreement for annealed specimens is offered.

The Luminescence of Diamonds, by C. BULL and G. F. J. GARLICK.

ABSTRACT. Studies of the luminescence characteristics of different types of industrial diamonds (80–100 mesh dust) have been made. Blue fluorescent diamonds (3,650 A. exciting radiation) exhibit thermoluminescence characteristics of two groups of metastable electronic states, with mean activation energies of 0·5 and 0·7 ev. respectively. Transitions of a forbidden nature from these states to the normal states of emission centres give rise to a temperature-independent, green-yellow phosphorescence, whose longest-lived component has a decay constant of 2×10^{-5} sec^{-1}. Most yellow luminescent and non-luminescent diamonds (3,650 A. exciting radiation) exhibit a blue luminescence when excited by energetic particles or by X-rays, but no thermoluminescence is observed after any type of excitation. The luminescence of diamonds appears to be characteristic of the matrix crystal, and a tentative model for the emission centres is given.